DATA CENTER

SAN Design and Best Practices

Version 2.3

A high-level guide focusing on Fibre Channel Storage Area Network (SAN) design and best practices, covering planning, topologies, device sharing in routed topologies, workload monitoring, and detecting server and storage latencies—to help with decisions required for successful SAN design.

BROCADE

CONTENTS

INTRODUCTION 6
Audience and Scope 6
Approach 6
Overview 6
ARCHITECTING A SAN 7
Operational Considerations 8
Be the Pilot 8
Diagnostics—Gen 5 Fibre Channel Platforms Only 8
SAN DESIGN BASICS 9
Topologies 9
Edge-Core Topology 10
Edge-Core-Edge Topology 10
Full-Mesh Topology 10
Redundancy and Resiliency 10
Switch Interconnections 11
ICL Connectivity for Brocade DCX and DCX 4-S Only 12
UltraScale ICL Connectivity for Brocade DCX 8510-8 and DCX 8510-4 with Gen 5 Fibre Channel Only 13
Brocade DCX 8510 UltraScale ICL Connection Best Practice 14
Mesh Topology 14
Device Placement 15
Traffic Locality 15
Fan-In Ratios and Oversubscription 17
DATA FLOW CONSIDERATIONS 20
Congestion in the Fabric 20
Traffic versus Frame Congestion 20
Sources of Congestion 20
Mitigating Congestion 21
Monitoring 21
Design Guidelines 22
AVAILABILITY AND HEALTH MONITORING 22
Health Monitoring 22
Brocade Fabric Watch 22
Brocade Fabric Watch Recommendations 23
RAS Log 23
Audit Log 23
Brocade SAN Health 23
Design Guidelines 23
Monitoring and Notifications 24
Available Paths 24
McDATA Interop Mode 24
Latencies 25

Misbehaving Devices 25
DESIGN GUIDELINES 26
Monitoring 26
IOPS and VMs 27
ROUTED TOPOLOGIES—METASANS 27
Backbone Considerations 29
Avoiding Congestion 31
Available Paths 31
Design Guidelines and Constraints for Routed SANs 31
VIRTUAL FABRICS TOPOLOGIES 32
VF Guidelines 32
Use Case: FICON and Open Systems (Intermix) 33
INTELLIGENT SERVICES 33
In-Flight Encryption and Compression—Gen 5 Fibre Channel Platforms Only 33
Virtual Fabric Considerations (Encryption and Compression) 34
In-Flight Encryption and Compression Guidelines 34
DISTANCE EXTENSION TOPOLOGIES 34
Buffer Allocation 34
Fabric Interconnectivity over Fibre Channel at Longer Distances 35
FC over IP (FCIP) 35
Basic FCIP Architectures 35
FCIP with FCR 38
Using EX_Ports and VEX_Ports 39
FCIP with FICON Environments 40
Advanced FCIP Configuration 41
IPsec 41
Compression 41
Adaptive Rate Limiting (ARL) 42
PerPriority-TCP-QoS 43
FCIP Design Best Practices 44
Bandwidth Allocation 44
FCIP Trunking 46
Protocol Optimization 47
Virtual Fabrics 48
Ethernet Interface Sharing 49
What is Not Supported? 50
WORKLOADS 51
Workload Virtualization 51
Intel-Based Virtualization Storage Access 51
Design Guidelines 52
Monitoring 52
Unix Virtualization 52
Recent Changes 52

Design Guidelines 53
Monitoring 54
SCALABILITY AND PERFORMANCE 54
SUPPORTABILITY 55
Firmware Upgrade Considerations 56
NPIV AND THE BROCADE ACCESS GATEWAY 56
Benefits of the Brocade Access Gateway 57
Constraints 58
Design Guidelines 58
Monitoring 58
Maintenance 59
BACKUP AND RESTORE 59
Determining SAN Bandwidth for Backups 59
Improving the Backup Infrastructure 60
STORAGE 61
Design Guidelines 61
Monitoring 61
Storage Virtualization 61
Design Guidelines 62
Monitoring 62
SECURITY 63
Zoning: Controlling Device Communication 63
Zone Management: Dynamic Fabric Provisioning (DFP) 63
Zone Management: Duplicate WWNs 63
Role-Based Access Controls (RBACs) 64
Access Control Lists (ACLs) 64
SCC Policy 64
DCC Policy 65
FCS Policy 65
IP Filter 65
Authentication Protocols 65
Policy Database Distribution 65
CAPACITY PLANNING 66
Gathering Requirements 66
Application Owners 66
Server and Storage Administrators 66
SAN Administrator: General 67
SAN Administrator: Backup and Restore 67
Facilities 68
Finance 68
TOOLS FOR GATHERING DATA 69
Brocade SAN Health 69
Power Calculator 69

Storage Traffic Patterns 69
Server Traffic Patterns 69
Backup Traffic Patterns 70
Tape Library 70
Backup Media Server 70
Brocade Network Advisor 70
SUMMARY 71
APPENDIX A: IMPORTANT TABLES 72
LWL Optics Support 72
APPENDIX B: MATRICES 73
Current Fabrics 73
Individual Fabric Details 73
Device Details 73
Metrics and Impact on SAN Design and Performance 74
Consolidated SAN Snapshot 75
Application-Specific Details 76
Quantitative Analysis: Radar Maps 77
SAN Admin Radar Map 77
Facilities Radar Map 78
APPENDIX C: PORT GROUPS 79
Brocade 5300 Trunk Port Groups 79
Brocade FC8-64 Trunk Groups 79
Gen 5 Fibre Channel Platforms 80
Brocade 6520 Trunk Groups 81
Brocade 6510 Trunk Groups 81
Brocade 6505 Trunk Groups 81
APPENDIX D: TERMINOLOGY 82
APPENDIX E: REFERENCES 83
Software and Hardware Product Documentation 83
Technical Briefs 83
Brocade Compatibility Matrix 83
Brocade Scalability Guidelines 83
Brocade SAN Health 83
Brocade FOS Features 83
Brocade Bookshelf 83
Other 83

INTRODUCTION

This document is a high-level design and best practices guide based on Brocade products and features, focusing on Fibre Channel SAN design. Covered topics include the early planning phase, understanding possible operational challenges, and monitoring and improving an already implemented SAN infrastructure.

Emphasis is given to in-depth information on how the data flows between devices affect a SAN infrastructure at a design and operational level.

The guidelines in this document do not apply to every environment but will help guide you through the decisions that you need to make for successful SAN design. Consult your Brocade representative or reference the documents in Appendix E for details about the hardware and software products, as well as the interoperability features mentioned in text and illustrations.

Note: *This is a "living" document that is continuously being expanded, so be sure to frequently check MyBrocade (my.brocade.com) for the latest updates of this and other best practice documents.*

Audience and Scope

This guide is for technical IT architects who are directly or indirectly responsible for SAN design based on the Brocade® 8 and Gen 5 Fibe Channel SAN platforms. It describes many of the challenges facing SAN designers today in both "greenfield" and legacy storage environments. While not intended as a definitive design document, this guide introduces concepts and guidelines to help you avoid potential issues that can result from poor design practices. This document describes best-practice guidelines in the following areas:

- Architecting a core data center infrastructure
- Capacity planning
- SAN topology
- Inter-switch connectivity
- Data flows
- Device connections
- Workloads/virtualization
- Distance extension
- Fibre Channel routing

Note: *A solid understanding of SAN concepts and Brocade Fibre Channel technology is assumed. Please refer to Appendix E for recommended additional publications.*

Approach

While some advanced features and specialized SAN applications are discussed, these topics are covered in greater detail in separate documents. The primary objective for this guide is to provide a solid foundation to facilitate successful SAN designs—designs that effectively meet current and future requirements. This document addresses basic administration and maintenance, including capabilities to identify early warning signs for end-device (initiator or target) latency, which can cause congestion in the SAN fabric. However, you should consult product documentation and documents in Appendix E for more details. A comprehensive discussion of SAN fabric administration best practices is covered in a separate document.

Overview

Although Brocade SAN fabrics are plug-and-play and can function properly even if left in a default state with ad hoc connections, Fibre Channel networks clearly benefit from a well thought out design and deployment strategy. In order to provide reliable and efficient delivery of data, your SAN topology should follow best practice guidelines based on SAN industry standards and considerations specific to Brocade.

This document does not consider physical environment factors such as power, cooling, and rack layout. Rather, the focus is on network connectivity (both inter-switch and edge device) and software configurations.

Note: The scope of this document is switch-centric and does not discuss end-device setup, configuration, as well as maintenance. Fabric monitoring, management, and diagnostics and McDATA and Brocade interoperability and migration are covered in separate documents.

ARCHITECTING A SAN

The SAN planning process is similar to any type of project planning and includes the following phases:

- **Phase I**—Gathering requirements
- **Phase II**—Developing technical specifications
- **Phase III**—Estimating project costs
- **Phase IV**—Analyzing Return on Investment (ROI) or Total Cost of Ownership (TCO) (if necessary)
- **Phase V**—Creating a detailed SAN design and implementation plan

When selecting which criteria to meet, you should engage users, server and storage Subject Matter Experts (SMEs), and other relevant experts to understand the role of the fabric. Since most SANs tend to operate for a long time before they are renewed, you should take future growth into account as SANs are difficult to re-architect. Deploying new SANs or expanding existing ones to meet additional workloads in the fabrics requires critical assessment of business and technology requirements. Proper focus on planning will ensure that the SAN, once it is deployed, meets all current and future business objectives, including availability, deployment simplicity, performance, future business growth, and cost. Tables in Appendix B are provided as a reference for documenting assets and metrics for SAN projects.

A critical aspect for successful implementation that is often overlooked is the ongoing management of the fabric. Identifying systems-level individual SMEs for all the components that make up the SAN, as well as adequate and up-to-date training on those components, is critical for efficient design and operational management of the fabric. When designing a new SAN or expanding an existing SAN, you should take into account these parameters:

Application virtualization

- Which applications will run under a Virtual Machine (VM) environment?
- How many VMs per server?
- Migration of VMs under what conditions (business and non-business hours, need additional CPU or memory to maintain response times)?
- Is there a need for SSDs to improve read response times?

Homogenous/heterogeneous server and storage platforms

- Do you use blade servers or rack servers?
- Is auto-tiering in place?
- Which Brocade Fabric OS® (FOS) runs in a multivendor environment?
- What is the refresh cycle of servers and storage platforms (2 years/3 years)?

Scalability

- How many user ports are needed now?
- How many Inter-Switch Links (ISLs)/Brocade UltraScale Inter-Chassis Links (ICLs) are required for minimizing congestion?
- Do you scale-out at the edge or the core?

Backup and disaster tolerance

- Is there a centralized backup? (This will determine the number of ISLs needed to minimize congestion at peak loads.)
- What is the impact of backup on latency-sensitive applications?
- Is the disaster solution based on a metro Fibre Channel (FC) or Fibre Channel over IP (FCIP) solution?

Diagnostics and manageability

- What is the primary management interface to the SAN (Command-Line Interface [CLI], Brocade Network Advisor, or third-party tool)?
- How often will Brocade FOS be updated?
- How will you validate cable and optics integrity?

Investment protection

- Support for future FC technology and interoperability
- Support for alternative technologies like FCoE (FC over Ethernet)

Operational Considerations

While Brocade fabrics are scalable in terms of port density and performance, the design goal should be to ensure simplicity for easier management, future expansion, and serviceability (for example, use edge-core-edge to a collapsed core; do not use both Inter-Fabric Routing IFR and Virtual Fabric (VF) if not really needed; turn on port monitoring and fencing parameters for critical application).

Note: Brocade has tested sharing of 64 devices per Logical SAN (LSAN) zone and 12 Fibre Channel Routers (FCRs) per backbone fabric with virtual fabrics. Any requirements beyond the tested configuration should be pre-tested in a non-production environment, or you should actively monitor system resources like CPU and memory utilization to minimize fabric anomalies.

Be the Pilot

Whether building a new SAN or connecting to an existing SAN, pre-staging and validating a fabric/application prior to putting it into production ensures there are baseline metrics in terms of rated throughput, latency, and expected Cyclic Redundancy Check (CRC) errors based on patch panel and physical cable infrastructure.

Diagnostics—Gen 5 Fibre Channel Platforms Only

For SANs deployed with Brocade Gen 5 Fibre Channel platforms, adapters, and optics, use the new ClearLink diagnostics port type, or D_Port, to run diagnostics during pre-deployment or when there are susceptible physical layer issues. Part of Brocade Fabric Vision technology, ClearLink is an offline diagnostics tool that allows users to automate a battery of tests to measure and validate latency and distance across the switch links. ClearLink diagnostics can also be used to verify the integrity of all 16 Gbps transceivers in the fabric on a one-by-one basis. In addition, a ClearLink diagnostic port requires only the individual ports that are attached to the link being tested to go offline, while leaving the rest of the ports to operate online, in isolation from the link. It can also be used to test links to a new fabric switch without allowing the new switch to join or even be aware of the current fabric. This fabric based physical layer validation enables the following:

- Local and long-distance measurements (5-meter granularity for 16 Gbps Small Form-factor Pluggables [SFPs] and 50 meters for 10 Gbps SFPs)
- Latency measurements
- Link performance
- Transceiver health check
- Transceiver uptime

The ClearLink diagnostic capability provides an opportunity to measure and thoroughly test ISLs before they are put into production. It is recommended that diagnostics be conducted prior to deployment or when there are CRC errors that could be caused by physical layer issues.

ClearLink D_Port guidelines and restrictions:

- Supported only on the Gen 5 Fibre Channel platforms with Brocade branded 10 Gbps or 16 Gbps transceivers
- Supported on E_Port and ports on both ends of the ISL must be in ClearLink D_Port mode
- Brocade FOS v7.1 provides ClearLink D_Port support on Gen 5 Fibre Channel switches running in Access Gateway mode as well as on links between Gen 5 Fibre Channel switches and Brocade fabric adapters running at 16 Gbps speed
- Brocade FOS v7.1 provides ClearLink D_Port support on UltraScale ICLs on the Brocade DCX® 8510 Backbone. The ClearLink D_Port on UltraScale ICLs skips the electrical and optical loopback tests, as the Quad Small Form-factor Pluggable (QSFP) used does not support it.
- If Brocade Inter-Switch Link (ISL) Trunking is deployed, use a minimum of 2 ports for the trunk. This enables the user to take down one of the links for diagnostic testing without disrupting the traffic on the remaining trunk members.
- Make sure there are at least two ISLs prior to taking a port offline for diagnostic testing. This ensures redundancy and prevents fabric segmentation in case a link is taken down for diagnostics.

SAN DESIGN BASICS

This section provides high-level guidelines necessary to implement a typical SAN installation. The focus is on best practices for core-edge or edge-core-edge fabrics. The discussion starts at the highest level, the data center, and works down to the port level, providing recommendations at each point along the way.

Topologies

A typical SAN design comprises devices on the edge of the network, switches in the core of the network, and the cabling that connects it all together. Topology is usually described in terms of how the switches are interconnected, such as ring, core-edge, and edge-core-edge or fully meshed. At this point the focus is on switch topology with SLs—device connectivity is discussed in later sections. The recommended SAN topology to optimize performance, management, and scalability is a tiered, core-edge topology (sometimes called core-edge or tiered core edge). This approach provides good performance without unnecessary interconnections. At a high level, the tiered topology has a large number of edge switches used for device connectivity, and a smaller number of core switches used for routing traffic between the edge switches, as shown in Figure 1.

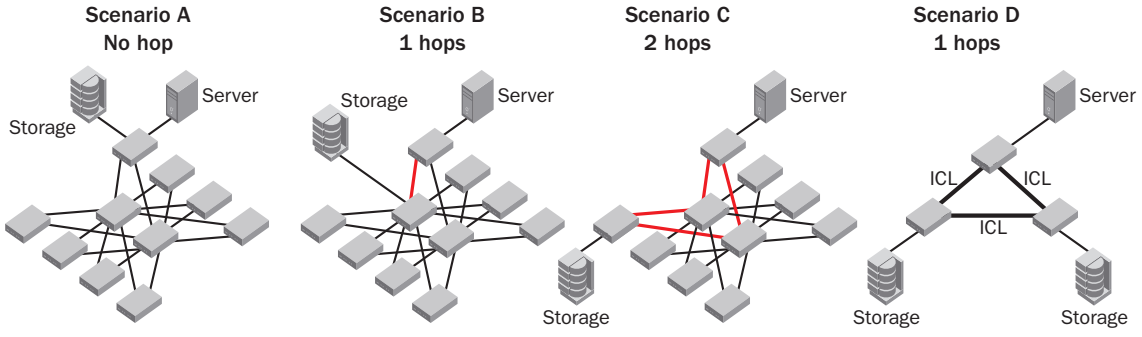

Figure 1. Four scenarios of tiered network topologies (hops shown in heavier, orange connections).

The difference between these four scenarios is device placement (where devices are attached to the network) and the associated traffic flow, which is discussed further in the "Data Flow Considerations" section later in this document.

- **Scenario A** has localized traffic, which can have small performance advantages but does not provide ease of scalability or manageability.
- **Scenario B**, also called edge-core, separates the storage and servers, thus providing ease of management and moderate scalability.
- **Scenario C** also known as edge-core-edge, has both storage and servers on edge switches, which provides ease of management and is much more scalable.
- **Scenario D** is a full-mesh topology, and server to storage is no more than one hop. Designing with UltraScale ICLs is an efficient way to save front-end ports, and users can easily build a large (for example, 1536-port or larger) fabric with minimal SAN design considerations.

Edge-Core Topology

The edge-core topology (Figure 1—Scenario B) places initiators (servers) on the edge tier and storage (targets) on the core tier. Since the servers and storage are on different switches, this topology provides ease of management as well as good performance, with most traffic only traversing one hop from the edge to the core. (Storage-to-storage traffic is two hops if the second storage is on another core switch], but the two cores can be connected if fabrics are redundant.) The disadvantage to this design is that the storage and core connections are in contention for expansion. In other words, this topology allows for only minimal growth.

Edge-Core-Edge Topology

The edge-core-edge topology (in Figure 1—Scenario C) places initiators on one edge tier and storage on another edge tier, leaving the core for switch interconnections or connecting devices with network-wide scope, such as Dense Wavelength Division Multiplexers (DWDMs), inter-fabric routers, storage virtualizers, tape libraries, and encryption engines. Since servers and storage are on different switches, this design enables independent scaling of compute and storage resources, ease of management, and optimal performance—with traffic traversing only two hops from the edge through the core to the other edge. In addition, it provides an easy path for expansion as ports and/or switches can readily be added to the appropriate tier as needed.

Full-Mesh Topology

A full-mesh topology (Figure 1—Scenario D) allows you to place servers and storage anywhere, since the communication between source to destination is no more than one hop. With optical UltraScale ICLs on the Brocade DCX 8510, customers can build a full-mesh topology that is scalable and cost effective compared to the previous generation of SAN products.

Note: *Hop count is not a concern if the total switching latency is less than the disk I/O timeout value.*

Redundancy and Resiliency

An important aspect of SAN topology is the resiliency and redundancy of the fabric. The main objective is to remove any single point of failure. Resiliency is the ability of the network to continue to function and/or recover from a failure, while redundancy describes duplication of components, even an entire fabric, to eliminate a single point of failure in the network. Brocade fabrics have resiliency built into Brocade FOS, the software that runs on all Brocade B-Series switches, which can quickly "repair" the network to overcome most failures. For example, when a link between switches fails, FSPF quickly recalculates all traffic flows. Of course this assumes that there is a second route, which is when redundancy in the fabric becomes important.

The key to high availability and enterprise-class installation is redundancy. By eliminating a single point of failure, business continuance can be provided through most foreseeable and even unforeseeable events. At the highest level of fabric design, the complete network should be redundant, with two completely separate fabrics that do not share any network equipment (routers or switches).

Servers and storage devices should be connected to both networks utilizing some form of Multi-Path I/O (MPIO) solution, such that data can flow across both networks seamlessly in either an active/active or active/passive mode. MPIO ensures that if one path fails, an alternative is readily available. Ideally, the networks would be identical, but at a minimum they should be based on the same switch architecture. In some cases, these networks are in the same location. However, in order to provide for Disaster Recovery (DR), two separate locations are often used, either for each complete network or for sections of each network. Regardless of the physical geography, there are two separate networks for complete redundancy.

In summary, recommendations for the SAN design are to ensure application availability and resiliency via the following:

- Redundancy built into fabrics to avoid a single point of failure
- Servers connected to storage via redundant fabrics
- MPIO-based failover from server to storage
- Redundant fabrics based on similar architectures
- Separate storage and server tiers for independent expansion
- At a minimum core switches should be of equal or higher performance compared to the edges.
- Define the highest performance switch in the fabric to be the principal switch.

Switch Interconnections

As mentioned previously, there should be at least two of every element in the SAN to provide redundancy and improve resiliency. The number of available ports and device locality (server/storage tiered design) determines the number of ISLs needed to meet performance requirements. This means that there should be a minimum of two trunks, with at least two ISLs per trunk. Each source switch should be connected to at least two other switches, and so on. In Figure 2, each of the connection lines represents at least two physical cable connections.

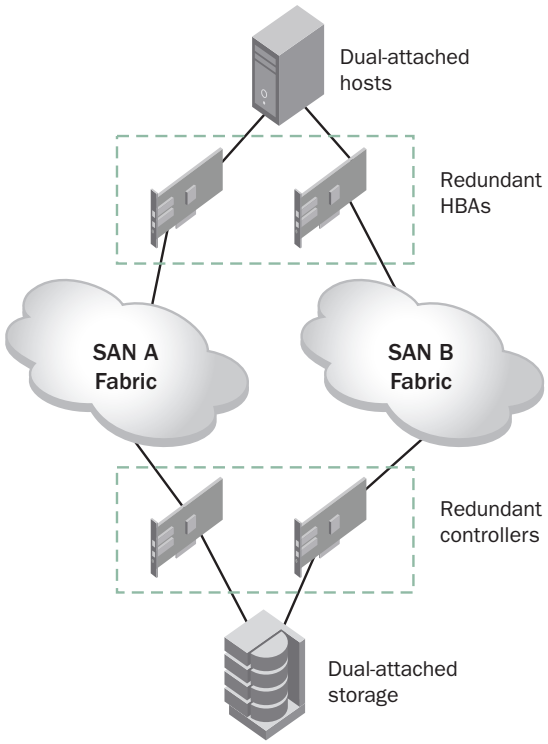

Figure 2. Connecting devices through redundant fabrics.

In addition to redundant fabrics, redundant links should be placed on different blades, different ASICs, or at least different port groups whenever possible, as shown in Figure 3. (Refer to Appendix C to determine trunk groups for various port blades. For more details, see the Brocade Fabric OS Administrator's Guide.) Whatever method is used, it is important to be consistent across the fabric. For example, do not place ISLs on lower port numbers in one chassis (as shown in the left diagram in Figure 3) and stagger them in another chassis (as shown in the right diagram in Figure 3). Doing so would be mismatched ISL placement.

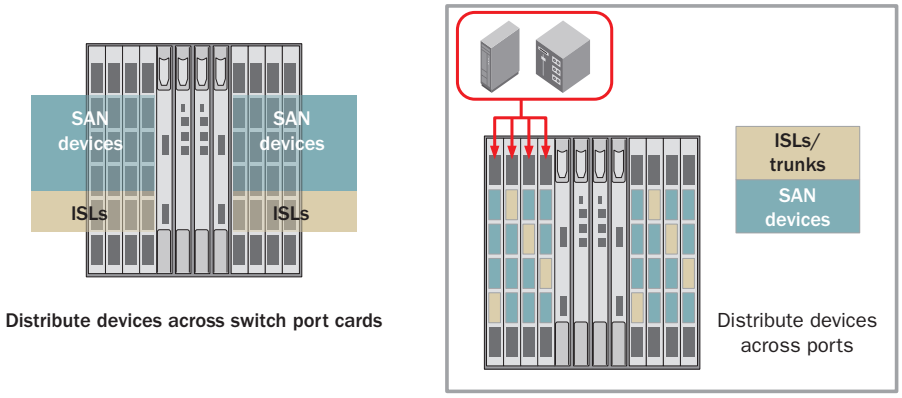

Figure 3. Examples of distributed ISL placement for redundancy.

Note: *In Figure 3, ISL trunks are placed on separate Application-Specific Integrated Circuits (ASICs) or port groups. It is important to match ISL placement between devices and across fabrics to ensure simplicity in design and assist in problem determination.*

ICL Connectivity for Brocade DCX and DCX 4-S Only

The Brocade DCX Backbone platform provides an additional method of interconnect called Inter-Chassis Links (ICLs). ICL ports are located on the core blades and provide 512 Gbps of bandwidth per chassis (equivalent to a 64-port blade) for additional inter-chassis connectivity. Two or three chassis can be interconnected (see Figures 4 and 5 for examples), freeing up front-end ports for connecting end devices.

A SAN topology should be evaluated for the feasibility of using ICLs between chassis to free up regular blade ports.

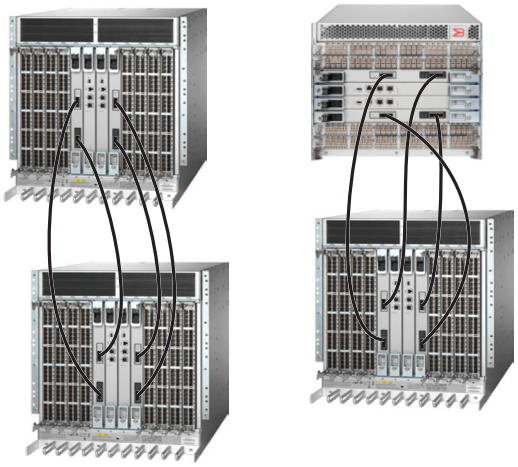

Figure 4. Examples of two-chassis ICL configurations: Brocade DCX to DCX (left) and Brocade DCX-4S Backbone to DCX (right).

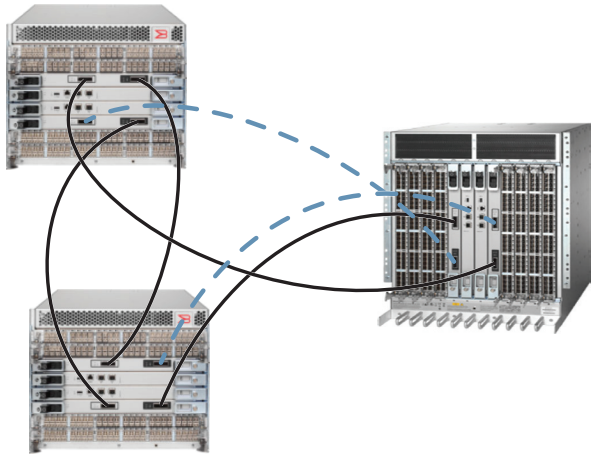

Figure 5. Example of three-chassis ICL configuration for 8 Gbps platform.

Note: *Refer to the Brocade DCX Hardware Reference Manual for detailed ICL connectivity. ICLs can be used instead of ISLs for a Brocade DCX/DCX-4S core-edge fabric—taking into account that the ICL cable length is 2 meters or less. Also, an ICL connection is considered a "hop of no concern" in a FICON environment.*

UltraScale ICL Connectivity for Brocade DCX 8510-8 and DCX 8510-4 with Gen 5 Fibre Channel Only

The Brocade DCX 8510-8 and DCX 8510-4 platforms use second-generation UltraScale ICL technology from Brocade with optical QSFP. The Brocade DCX 8510-8 allows up to 32 QSFP ports, and the Brocade DCX 8510-4 allows up to 16 QSFP ports to help preserve switch ports for end devices. Each QSFP port actually has four independent 16 Gbps links, each of which terminates on a different ASIC within the core blade. Each core blade has four ASICs. A pair of connections between two QSFP ports can create 32 Gbps of bandwidth. Figure 6 shows a core-edge design based on UltraScale ICLs supporting 2304 16 Gbps ports with a minimum of 256 Gbps of bandwidth between the chassis (12:1 oversubscription). As more UltraScale ICLs are added, oversubscription can be reduced to 6:1.

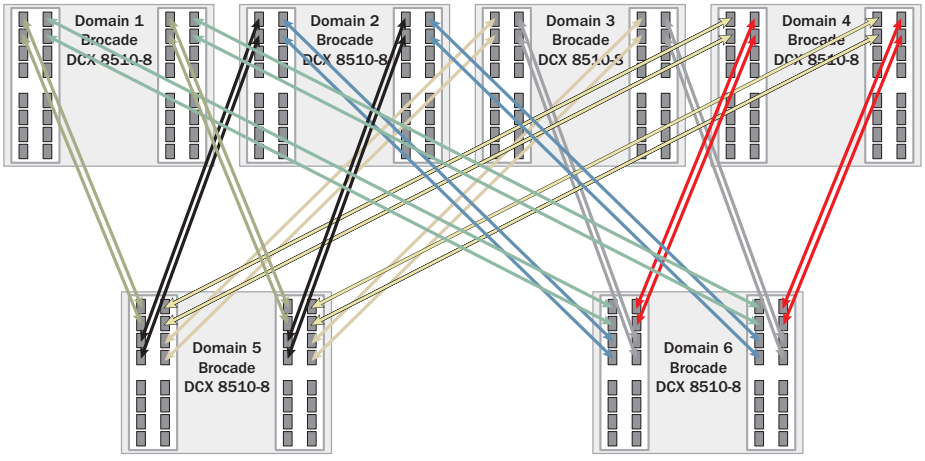

Figure 6. UltraScale ICL-based core-edge design.

To connect multiple Brocade DCX 8510 chassis via UltraScale ICLs, a minimum of four ICL ports (two on each core blade) must be connected between each chassis pair, as shown in Figure 7. With 32 ICL ports available on the Brocade DCX 8510-8 (with both ICL POD licenses installed), this supports ICL connectivity with up to eight other chassis and at least 256 Gbps of bandwidth to each connected Brocade DCX 8510. The dual connections

on each core blade must reside within the same ICL trunk boundary on the core blades. If more than four ICL connections are required between a pair of Brocade DCX 8510 chassis, additional ICL connections should be added in pairs (one on each core blade).

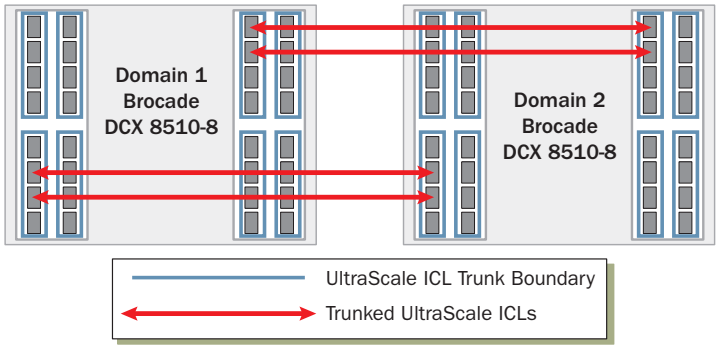

Figure 7. Minimum connections needed between Brocade DCX 8510 chassis.

Brocade DCX 8510 UltraScale ICL Connection Best Practice

Each core blade in a chassis must be connected to each of the two core blades in the destination chassis to achieve full redundancy. (Note: For redundancy, use at least one pair of links between 2 core blades.)

Mesh Topology

A mesh design provides a single hop between source and destination, and initial Brocade FOS v7.0 release supports a 3 chassis mesh design (same as existing 8 Gb platform) with 15/25/50 meter distance. Brocade FOS v7.1 provides support for 100 meters with select QSFPs and OM4 fiber. In the configuration shown in Figure 8, up to 1152 16 Gbps ports are supported using UltraScale ICLs with a 12:1 oversubscription. As more UltraScale ICLs are added, oversubscription can be reduced to 3:1.

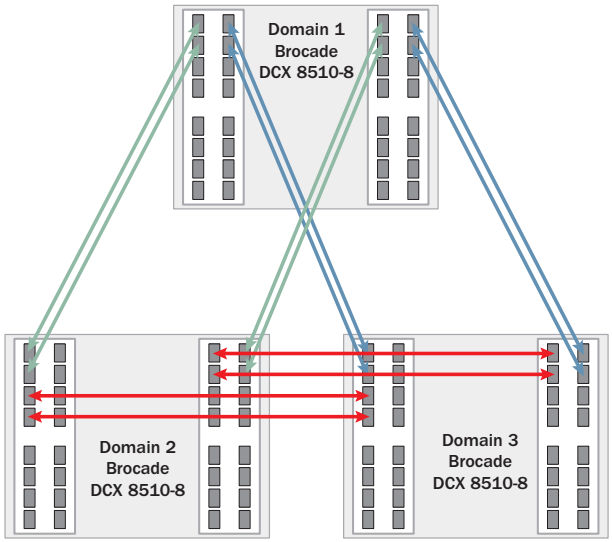

Figure 8. UltraScale ICL-based full-mesh topology.

NOTE: *Refer to the Hyper-Scale Fabrics: Scale-out Architecture with Brocade DCX 8510 Feature Brief for details. UltraScale ICL connections are considered a "hop of no concern" in a FICON fabric.*

Brocade recommends core-edge or edge-core-edge as the primary SAN design methodology, or mesh topologies used for small fabrics (under 2000 ports). As a SAN design best practice, edge switches should connect to at least two core switches with trunks of at least two ISLs each. Each of those trunks should be attached to a different blade/port group. In order to be completely redundant, there would be a completely mirrored second fabric and devices need to be connected to both fabrics, utilizing MPIO.

Recommendations for switch ISL/UltraScale ICL connectivity are:

- There should be at least two core switches.
- Every edge switch should have at least two trunks to each core switch.
- Select small trunk groups (keep trunks to two ISLs) unless you anticipate very high traffic volumes. This ensures that you can lose a trunk member without losing ISL connectivity.
- Place redundant links on separate blades.
- Trunks should be in a port group (ports within an ASIC boundary).
- Allow no more than 30m in cable difference for optimal performance for ISL trunks.
- Use the same cable length for all UltraScale ICL connections.
- Avoid using ISLs to the same domain if there are UltraScale ICL connections.
- Use the same type of optics on both sides of the trunks: Short Wavelength (SWL), Long Wavelength (LWL), or Extended Long Wavelength (ELWL).

Device Placement

Device placement is a balance between traffic isolation, scalability, manageability and serviceability. With the growth of virtualization and multinode clustering on the UNIX platform, frame congestion can become a serious concern in the fabric if there are interoperability issues with the end devices.

Traffic Locality

Designing device connectivity depends a great deal on the expected data flow between devices. For simplicity, communicating hosts and targets can be attached to the same switch.

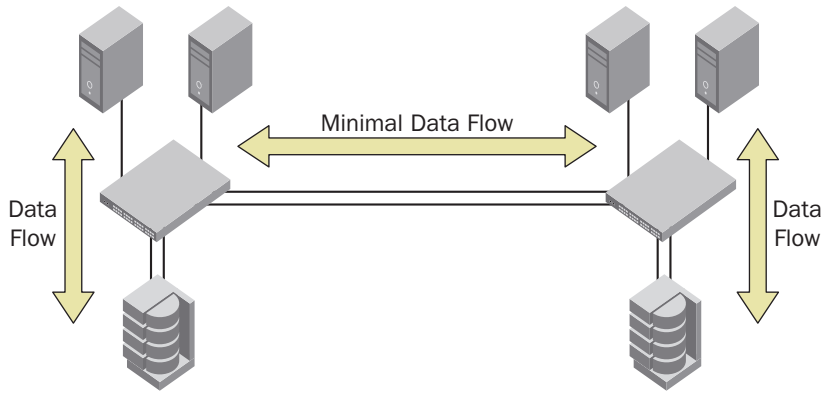

Figure 9. Hosts and targets attached to the same switch to maximize locality of data flow.

However, this approach does not scale well. Given the high-speed, low-latency nature of Fibre Channel, attaching these host-target pairs on different switches does not mean that performance is adversely impacted. Though traffic congestion is possible, it can be mitigated with proper provisioning of ISLs/UltraScale ICLs. With current generation switches, locality is not required for performance or to reduce latencies. For mission-critical applications, architects may want to localize the traffic when using Solid State Drives (SSDs) or in very exceptional cases, particularly if the number of ISLs available is restricted or there is a concern for resiliency in a multi-hop environment.

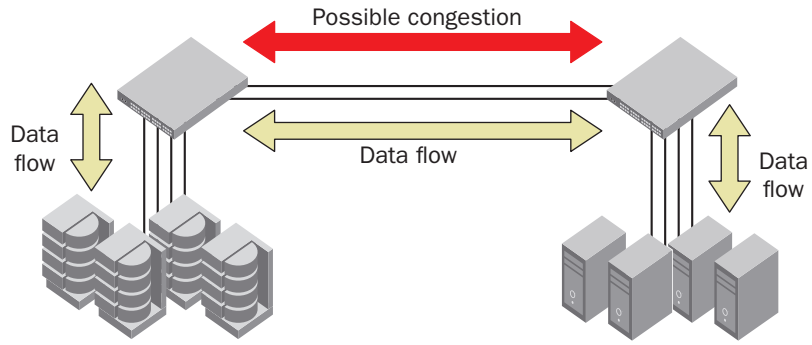

Figure 10. Hosts and targets attached to different switches for ease of management and expansion.

One common scheme for scaling a core-edge topology is dividing the edge switches into a storage tier and a host/initiator tier. This approach lends itself to ease of management as well as ease of expansion. In addition, host and storage devices generally have different performance requirements, cost structures, and other factors that can be readily accommodated by placing initiators and targets in different tiers.

The following topology provides a clearer distinction between the functional tiers.

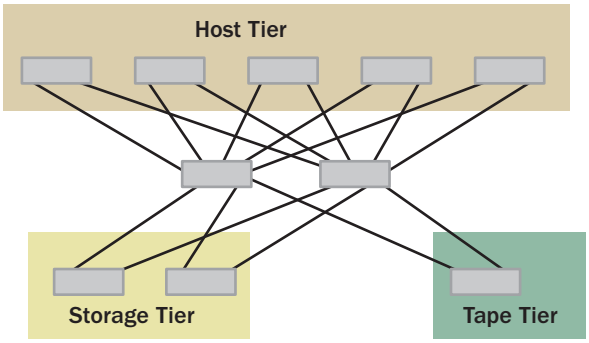

Figure 11. Device type based edge-core-edge tiered topology.

Recommendations for device placement include:

- The best practice fabric topology is core-edge or edge-core-edge with tiered device connectivity, or full-mesh if the port count is less than 2000 ports.
- Minimize the use of localized traffic patterns and, if possible, keep servers and storage connected to separate switches.
- Select the appropriate optics (SWL/LWL/ELWL) to support the distance between switches and devices and switches.

Fan-In Ratios and Oversubscription

Another aspect of data flow is the "fan-in-ratio" or "oversubscription", in terms of source ports to target ports and device to ISL's. This is also referred to as the "fan-out-ratio" if viewed from the storage array perspective. The ratio is the number of device ports that share a single port, whether ISL, UltraScale ICL, or target. This is always expressed from the single entity point of view, such as 7:1 for 7 hosts utilizing a single ISL or storage port.

What is the optimum number of hosts that should connect per to a storage port? This seems like a fairly simple question. However, once you take into consideration clustered hosts, VMs, and number of Logical Unit Numbers (LUNs) (storage) per server the situation can quickly become much more complex. Determining how many hosts to connect to a particular storage port can be narrowed down to three considerations: port queue depth, I/O per second (IOPS), and throughput. Of these three, throughput is the only network component. Thus, a simple calculation is to add up the expected bandwidth usage for each host accessing the storage port. The total should not exceed the supported bandwidth of the target port, as shown in Figure 12.

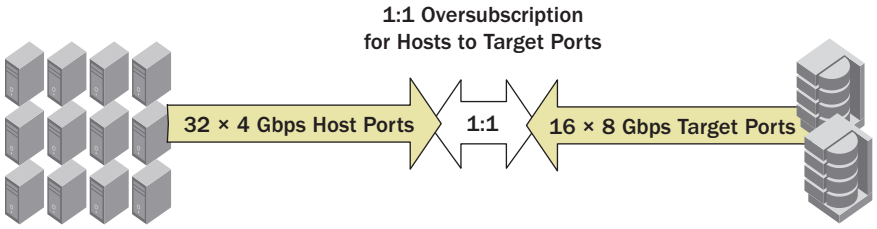

Figure 12. Example of one-to-one oversubscription.

In practice, however, it is highly unlikely that all hosts perform at their maximum level at any one time. With the traditional application-per-server deployment, the Host Bus Adapter (HBA) bandwidth is overprovisioned. However, with virtual servers (KVM, Xen, Hyper-V, proprietary Unix OSs, and VMware) the game can change radically. Network oversubscription is built into the virtual server concept. To the extent that servers leverage virtualization technologies, you should reduce network-based oversubscription proportionally. It may therefore be prudent to oversubscribe ports to ensure a balance between cost and performance. An example of 3 to 1 oversubscription is shown in Figure 13.

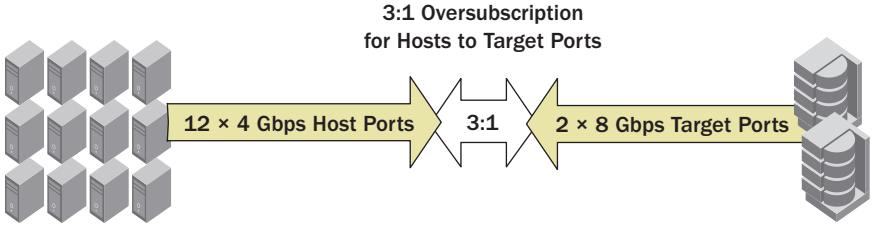

Figure 13. Example of three-to-one oversubscription.

Another method is to assign host ports to storage ports based on capacity. The intended result is a small number of high-capacity hosts and a larger number of low-capacity servers assigned to each storage port, thus distributing the load across multiple storage ports. Figure 14 shows the impact of the two different LUN provisioning strategies described above. Notice that there is a huge diffenence between the fan-in to the storage port, based on the number of LUNs provisioned behind the port.

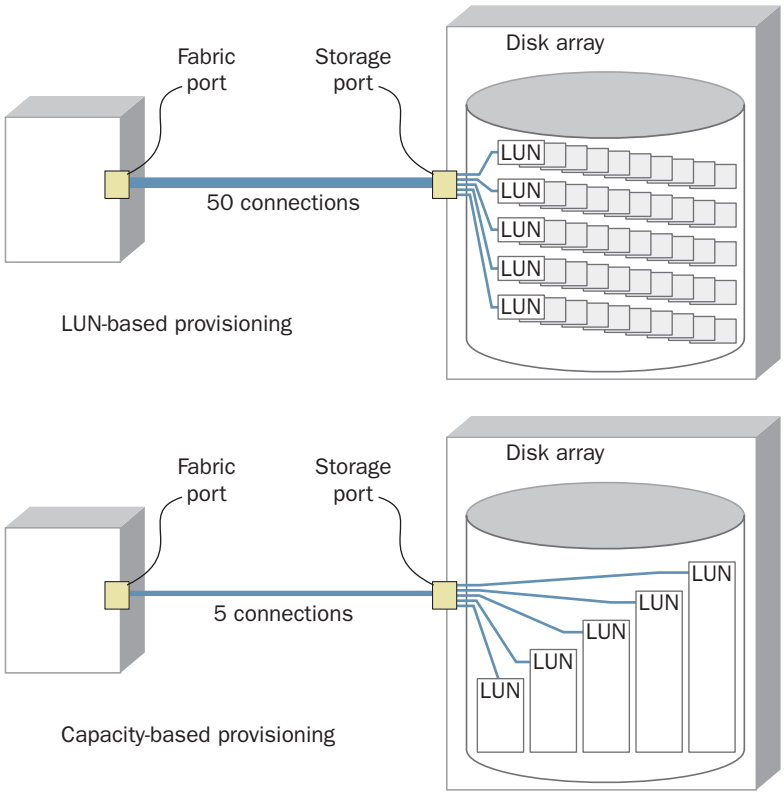

Figure 14. Two different LUN provisioning strategies.

Regardless of the method used to determine the fan-In/fan-out ratios, port monitoring should be used to determine actual utilization and what adjustments, if any, should be made. In addition, ongoing monitoring provides useful heuristic data for effective expansion and efficient assignment of existing storage ports. For determining the device-to-ISL fan-in ratio, a simple calculation method works best: the storage port should not be oversubscribed into the core (Example: an 8 Gbps storage port should have an 8 Gbps pipe into the core).

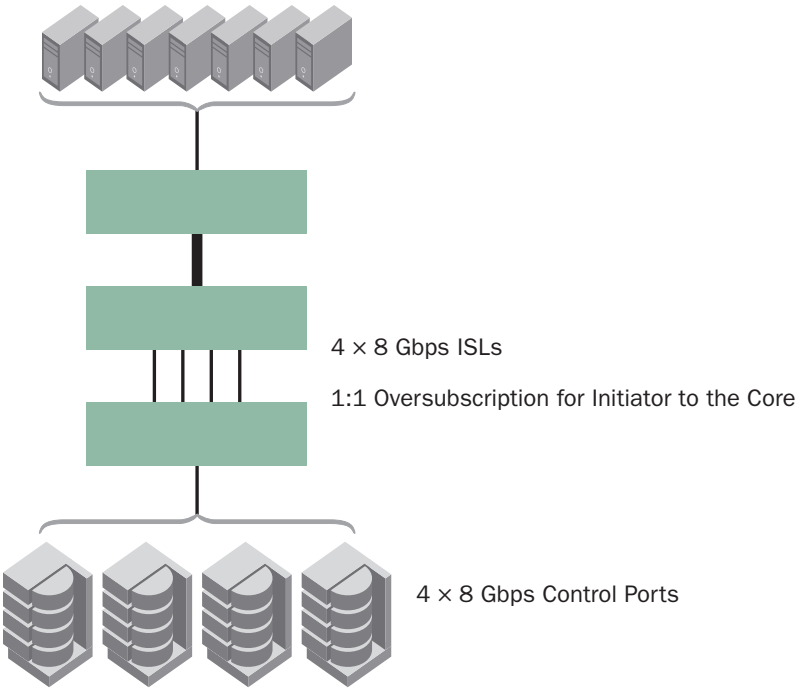

Figure 15. One-to-one oversubscription for targets into the core.

The realized oversubscription ratio of host-to-ISL should be roughly the same as the host-to-target ratio, taking into account the bandwidth (that is, if there are four hosts accessing a single 4 Gbps storage port, then those four hosts should have a 4 Gbps pipe into the core.) In other words, match device utilization and speeds with ISL speeds, as shown in Figure 16.

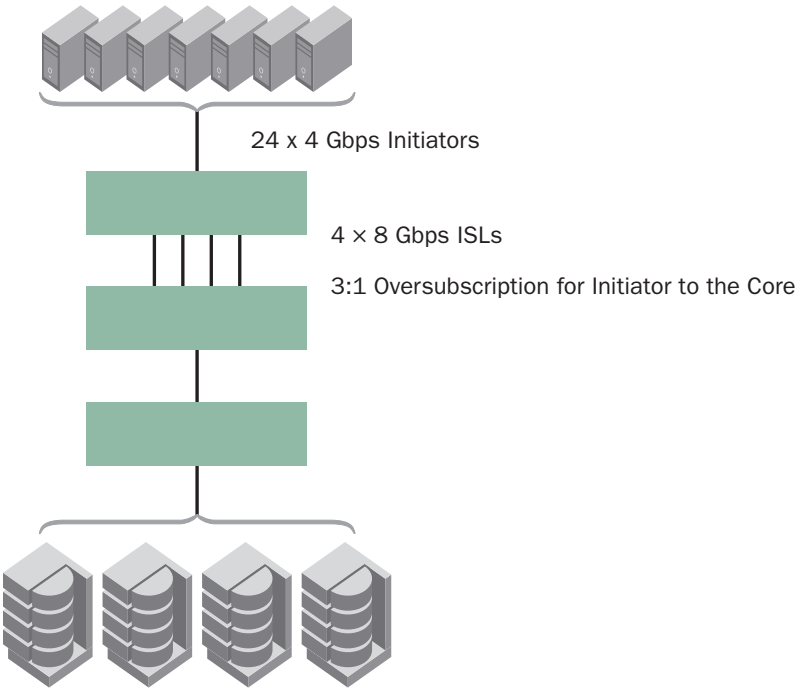

Figure 16. Three-to-one oversubscription for hosts coming into the core.

Recommendations for avoiding frame congestion (when the number of frames is the issue rather than bandwidth utilization) include:

- Use more and smaller trunks.
- Storage ports should follow the array vendor-suggested fan-in ratio for ISLs into the core. Follow vendor-suggested recommendations when implementing a large number of low-capacity LUNs.
- Bandwidth through the core (path from source/host to destination/target) should exceed storage requirements.
- Host-to-core subscription ratios should be based on both the application needs and the importance of the application.
- Plan for peaks, not average usage.
- For mission-critical applications, the ratio should exceed peak load enough such that path failures do not adversely impact the application. In other words, have enough extra bandwidth to avoid congestion if a link fails.

DATA FLOW CONSIDERATIONS

Congestion in the Fabric

Congestion is a major source of poor performance in a fabric. Sufficiently impeded traffic translates directly into poor application performance.

There are two major types of congestion: traffic-based and frame-based. Traffic-based congestion occurs when link throughput capacity is reached or exceeded and the link is no longer able to pass more frames. Frame-based congestion occurs when a link has run out of buffer credits and is waiting for buffers to free up to continue transmitting frames.

Traffic versus Frame Congestion

Once link speeds reach 4 Gbps and beyond, the emphasis on fabric and application performance shifts from traffic -level issues to frame congestion. It is very difficult with current link speeds and Brocade features such as Brocade ISL Trunking or UltraScale ICLs to consistently saturate a link. Most infrastructures today rarely see even two-member trunks reaching a sustained 100 percent utilization. Frame congestion can occur when the buffers available on a Fibre Channel port are not sufficient to support the number of frames the connected devices wish to transmit. This situation can result in credit starvation backing up across the fabric. This condition is called back pressure, and it can cause severe performance problems.

One side effect of frame congestion can be very large buffer credit zero counts on ISLs and F_Ports. This is not necessarily a concern, unless counts increase rapidly in a very short period of time. Brocade has added a new feature, Bottleneck Detection, to more accurately assess the impact of a lack of buffer credits.

The sources and mitigation for traffic are well known and are discussed at length in other parts of this document. The remainder of this section focuses on the sources and mitigation of frame-based congestion.

Sources of Congestion

Frame congestion is primarily caused by latencies somewhere in the SAN—usually storage devices and occasionally hosts. These latencies cause frames to be held in ASICs and reduce the number of buffer credits available to all flows traversing that ASIC. The congestion backs up from the source of the latency to the other side of the connection and starts clogging up the fabric. Back pressure can be created from the original source of the latency to the other side and all the way back (through other possible paths across the fabric, to the original source again. Once this situation arises the fabric is very vulnerable to severe performance problems.

Sources of high latencies include:

- Storage devices that are not optimized or where performance has deteriorated over time
- Distance links where the number of allocated buffers has been miscalculated or where the average frame sizes of the flows traversing the links has changed over time
- Hosts where the application performance has deteriorated to the point that the host can no longer respond to incoming frames in a sufficiently timely manner
- Incorrectly configured HBA's
- Massive oversubscription on target ports and ISL's
- Tape devices

Other contributors to frame congestion include behaviors where short frames are generated in large numbers such as:

- Clustering software that verifies the integrity of attached storage
- Clustering software that uses control techniques such as SCSI RESERVE/RELEASE to serialize access to shared file systems
- Host-based mirroring software that routinely sends SCSI control frames for mirror integrity checks
- Virtualizing environments, both workload and storage, that use in-band Fibre Channel for other control purposes

Mitigating Congestion

Frame congestion cannot be corrected in the fabric. Devices exhibiting high latencies, whether servers or storage arrays, must be examined and the source of poor performance eliminated. Since these are the major sources of frame congestion, eliminating them typically addresses the vast majority of cases of frame congestion in fabrics.

Brocade has introduced a new control mechanism in an attempt to minimize the effect of some latencies in the fabric. Edge Hold Time (EHT) is a new timeout value that can cause some blocked frames to be discarded earlier by an ASIC in an edge switch where the devices typically are provisioned. EHT is available from Brocade FOS v6.3.1b or later and allows for frame drops for shorter timeout intervals than the 500 milliseconds typically defined in the Fibre Channel Standard. EHT accepts values from 500 all the way down to 80 milliseconds. The EHT default setting for F_Ports is 220 milliseconds and the default EHT setting for E_Ports is 500 milliseconds. Note that an I/O retry is required for each of the dropped frames, so this solution does not completely address high-latency device issues.

EHT applies to all the F_Ports on a switch and all the E_Ports that share the same ASIC as F_Ports. It is a good practice to place servers and ISLs on different ASICs since the EHT value applies to the entire ASIC, and it is recommended that the ISL EHT stay at 500 ms.

Note: *EHT applies to the switch and is activated on any ASIC that contains a F_Port. (For example, if EHT is set to 250ms and the ASIC contains F_Ports and E_Ports, the timeout value for all the ports is 250 ms.*

Behaviors that generate frequent large numbers of short frames cannot typically be changed—they are part of the standard behavior of some fabric-based applications or products. As long as the major latencies are controlled, fabrics tolerate this behavior well.

Monitoring

A recent Brocade FOS feature, Bottleneck Detection, was introduced to directly identify device and link latencies and high link utilization.

Bottleneck Detection, when applied to F_Ports (devices) detects high-latency devices and provides notification on the nature and duration of the latency. This is a huge advantage to the storage administrator, because there is now a centralized facility that can potentially detect storage latencies while they are still intermittent.

Bottleneck detection can also serve as a confirmation to host information when storage latencies are suspected in poor host performance. The reverse (eliminating the storage as the source of poor performance) is also true.

Beginning with Brocade FOS v6.4, Bottleneck Detection can also be applied to ISLs (E_Ports) and will highlight issues on those links.

The sampling interval and number of notifications are configurable, as well as the alerting mechanisms. With Brocade FOS v6.4 notifications can be configured for Reliability, Availability, and Serviceability (RAS) log and Simple Network Management Protocol (SNMP). Brocade Network Advisor can be configured to automatically monitor and detect bottlenecks in the fabric. You can easily pinpoint areas of network congestion with visual connectivity maps and product trees.

Design Guidelines

Edge Hold Time (EHT):

- Recommended primarily for initiators (hosts). Extreme care must be taken if you choose to apply EHT to target ports because a target port can service a large number of initiators. A large number of frame drops on a target port can potentially affect a very large number of running applications. Those applications may be more tolerant to poor performance that to a large number of I/O retries.
- There is no calculation for determining the best value for EHT. EHT can be set from 100 to 500 milliseconds. The lower the value, the more frame drops you can expect. Brocade recommends that you take a value of approximately 250 milliseconds and observe the results.
- EHT is less effective when initiators and targets share the same switch, because the timeout value will apply equally to both storage and host ports.
- EHT applies to the entire ASIC. If possible, ISLs should be placed on a different ASIC than the servers.

Bottleneck Detection:

- A phased approach to deploying Bottleneck Detection works best. Given the potential for a large number of alerts early on in the process, Brocade recommends starting with a limited number of storage ports and incrementally increasing the number of ports monitored over time. Once the storage latencies are dealt with, you should move on to the host (initiator) ports and ISLs. You can increase the number of ports monitored once the chronic latency problems have been dealt with.
- Bottleneck Detection consumes some switch memory to keep some historical data. Brocade recommends no more than 100 ports in total be monitored at once on Brocade 48000 platforms, to avoid any potential for memory issues. There are no such limitations on the Brocade DCX 8510 with Gen 5 Fibre Channel or DCX platforms.

AVAILABILITY AND HEALTH MONITORING

With Brocade FOS and Brocade Network Advisor, IT organizations can monitor fabrics on both a real-time and historical basis. This allows users to address performance issues proactively and rapidly diagnose the underlying causes, then quickly resolve the issues before the SAN becomes the bottleneck for critical applications. An overview of the major components is provided below. A complete guide to health monitoring is beyond the scope of this document. Please refer to the Brocade Fabric OS Command Reference Guide, the Brocade Fabric OS Troubleshooting Guide, the appropriate Brocade SAN Health and Fabric Watch guides, and the Brocade Network Advisor SAN User Manual for more detailed information.

Health Monitoring

Brocade Fabric Watch

Fabric Watch is an optional health monitor that allows you to constantly monitor each director or switch for potential faults and automatically alerts you to problems long before they become costly failures.

Fabric Watch tracks a variety of SAN fabric elements and events. Monitoring fabric-wide events, ports, and environmental parameters enables early fault detection and isolation as well as performance measurement. You can configure fabric elements and alert thresholds on an individual port basis, and you can also easily integrate Fabric Watch with enterprise system management solutions.

Fabric Watch provides customizable monitoring thresholds. You can configure Fabric Watch to provide notification before problems arise, such as reporting when network traffic through a port is approaching the bandwidth limit. This information enables you to perform pre-emptive network maintenance, such as trunking or zoning, and avoid potential network failures.

Fabric Watch lets you define how often to measure each switch and fabric element and specify notification thresholds. Whenever fabric elements exceed these thresholds, Fabric Watch automatically provides notification using several methods, including e-mail messages, SNMP traps, and log entries.

Fabric Watch was significantly upgraded starting in Brocade FOS v6.4, and it continues to be a major source of early warning for fabric issues. Useful enhancements, such as port fencing to protect the fabric against misbehaving devices, are added with each new release of Brocade FOS.

Brocade Fabric Watch Recommendations

Brocade Fabric Watch is an optional feature that provides monitoring of various switch elements. Brocade Fabric Watch monitors ports based on the port type, for example, F_Port and E_Port classes, without distinguishing between initiators and targets. Since the monitoring thresholds and desired actions are generally different for initiators and targets, it is recommended that these devices be placed on different switches so that Brocade Fabric Watch settings can be applied accordingly.

Note: *For additional details, see the Brocade Fabric Watch Administrator's Guide.*

RAS Log

RAS log is the Brocade FOS error message log. Messages are organized by Brocade FOS component, and each one has a unique identifier as well as severity, source and platform information and a text message.

RAS log is available from each switch and director via the "errdump" command. RAS log messages can be forwarded to a syslog server for centralized collection or viewed within Brocade Network Advisor via the Master Log.

Audit Log

The Audit log is a collection of information created when specific events are identified on a Brocade platform. The log can be dumped via the auditdump command, and audit data can also be forwarded to a syslog server for centralized collection.

Information is collected on many different events associated with zoning, security, trunking, FCIP, FICON, and others. Each release of the Brocade FOS provides more audit information.

Brocade SAN Health

Brocade SAN Health provides snapshots of fabrics showing information such as switch and firmware levels, connected device information, snapshots of performance information, zone analysis, and ISL fan-in ratios.

Design Guidelines

Brocade strongly recommends implementing some form of monitoring of each switch. Often issues start out relatively benignly and gradually degrade into more serious problems. Monitoring the logs for serious and error severity messages will go a long way in avoiding many problems.

- Plan for a centralized collection of RAS log, and perhaps Audit log, via syslog. You can optionally filter these messages relatively easily through some simple Perl programs.

- Brocade platforms are capable of generating SNMP traps for most error conditions. Consider implementing some sort of alerting mechanism via SNMP.

Monitoring and Notifications

Error logs should be looked at regularly. Many end users use combinations of syslog and SNMP with the Brocade Fabric Watch and the logs to maintain a very close eye on the health of their fabrics. You can troubleshoot network-related issues such as syslog events and SNMP traps through the Event Manager within Brocade Network Advisor.

Brocade Network Advisor also collects, monitors and graphically displays real-time and historical performance data, so you can proactively manage your SAN network.

Brocade Professional Services can be engaged to assist with implementing these and other advanced features.

Available Paths

It is recommended that the SAN be deployed with at least two paths between source and destination.

Often, there are more than two paths and the utilization of these paths is dependent on the routing policy configuration.

- **Port-Based Routing** (PBR) assigns a single route between source port and destination port. Although this minimizes disruption caused by changes in the fabric, it represents a less efficient use of available bandwidth.
- **Exchange-Based Routing** (EBR) uses all available (equal-cost) routes between source port and destination port, with individual exchanges assigned a single route. Although it represents a more efficient use of available bandwidth, it is potentially more disruptive unless Dynamic Load Sharing (DLS) is implemented with the lossless feature.

The number of available paths can be adjusted by changing the size of trunk groups. While a trunk can have two to eight members, it may prove beneficial to have more trunks with fewer members. Spreading ISLs across multiple trunks uses more of the fabric bandwidth by spreading traffic across more paths. Keep at least two members in each trunk to avoid unnecessary frame loss if a trunk member fails.

McDATA Interop Mode

A mixed fabric is one with Brocade FOS switches and McDATA Enterprise OS switches (M-EOS). The inter-operability mode of the switch with McDATA switches can be McDATA Fabric mode, or McDATA Open Fabric mode. (Refer to the Brocade FOS Release Notes for supported platforms and restrictions.) McDATA Open Fabric mode is intended specifically for adding Brocade FOS-based products to M-EOS fabrics that are already using Open Fabric mode.

Brocade default routing protocol is EBR. McDATA default is PBR. Since these protocols operate very differently in how traffic is distributed across the ISLs, the resulting lopsided flow control for ISLs causes performance issues. McDATA's "open trunking" can be used to assist with the reallocation of flows to better ISLs; however, it will not be very effective against short duration flows (5–10 seconds) of small frames. This "micro-burst" sustained over time can result in fabric-wide performance problems due to frame congestion. So, to mitigate this possibility, or if you are experiencing this, you should try to identify these "heavy hitter" servers and when (what time of day) the micro-bursts are happening. You will be able to correlate the performance spike with errors on other non-related flows (collateral damage) from the non-discriminating C3 Discards on the ISLs they are sharing. (From an M-Series [legacy McDATA] standpoint, a single ISL is used.)

Here are some recommendations, listed in order of their level of invasiveness. Some customers are willing to move a server, but some are not. Also, some of these solutions will not work in certain environments, so begin planning the appropriate solution as early as possible:

Solution 1: Make the B-Series act more like the M-Series by using PBR versus EBR on the B-Series so that it will have the same single flow per ISL, just like the M-Series.

Solution 2: Migrate sooner to a B-Series core and B-Series edge from B-Series core and M-Series edge (or vice versa). Most customers find themselves in Interop Fabrics temporarily.

Solution 3: Move the heavy hitters (microbursts) to the core (in a core-edge topology). This will reduce the traffic on the ISLs and reduce the chances of frame congestion.

Solution 4: Install "Open Trunking" on M-Series to assist with load balancing ISLs based on flow and increase BBCs (Buffer to Buffer Credits) on a B-Series core switch. This will allow the M-Series to send more frames upward on the single ISL it has chosen, versus the B-Series flows downward will use ALL ISLs (EBR) and essentially equalize the flows from a BBC standpoint.

Example: Assume that a B-Series has 3 ISLs to M-Series. In this design, each B-Series will have 8 BBCs by default and the M-Series will have 16 by default. In this example, the B-Series will see a total of 24 BBCs toward the McDATA, whereas the McDATA will only see 16 and upward. This may not be enough BBCs on the single ISL to sustain the flow. The recommendation would be to assign the B-Series with 48 BBCs to equalize. The M-Series originated flows will now be able to fully utilize the ISL with 48 BBCs from the B-Series, whereas the M-Series will receive from the B-Series 48 total frames, as the B-Series will use all 3 ISLs to forward frames (EBR), since each M-Series is set to 16 BBCs by default. Keep in mind, the BBC value is really a receiver buffer setting.

Note: *Refer to the Brocade FOS Release Notes for McDATA interop support. Brocade FOS 7.0 and above do not support interoperability with the McDATA platform.*

Latencies

There are many causes of latencies:

- Slow devices such as storage arrays
- Oversubscribed devices
- Long-distance links
- Servers that are not responding rapidly enough to I/O requests they have previously made
- Degraded cables and SFPs causing many retried I/Os

There is very little that can be done in the fabric to accommodate end-device latencies: they typically must be addressed through other means. Array latencies can be dealt with by array or LUN reconfiguration or data migration. Long-distance problems might require more long-distance bandwidth or reconfiguration of the distance setting on the switch. Applications might require tuning to improve their performance, and failing links and SFPs must be identified and replaced. At best, the fabric can help identify the source of the problem. Brocade has been working hard to enhance RAS features in Brocade FOS in line with changing customer requirements. Some of these features are described briefly in the sections that follow.

Misbehaving Devices

All fabrics, regardless of the equipment vendor, are vulnerable to the effects of badly behaving devices, that is, a server or storage device that for some reason stops functioning or starts flooding the fabric with data or control frames. The effects of such behavior can be very severe, causing other applications to failover or even stop completely. There is nothing that the fabric can do to anticipate this behavior. Brocade has implemented several new features that are designed to rapidly detect a misbehaving device and isolate it from the rest of the fabric.

Isolating a single server has much less impact on applications than disabling a storage array port.

Typically, a storage port services many applications, and the loss of that storage can severely impact all the applications connected to it. One of the advantages of a core-edge design is that it is very simple to isolate servers from their storage and ensure that any action applied to host port for a given behavior can be very different than the action applied to a storage port for the same behavior.

DESIGN GUIDELINES

- **Transaction-based systems:** Make sure that ISL/UltraScale ICLs traversed by these systems to access their storage do not contain too many flows. The fan-in from the hosts/initiators should not exceed a ratio of 10 to 1. Also ensure that there is as little interference from other applications as possible, to ensure that latencies and congestion from other sources do not affect the overall performance of the applications.
- **I/O-intensive applications:** Bandwidth is the typical constraint for these systems. Modern fabrics typically provide more bandwidth than is needed, except for the most powerful hosts. Take care to ensure that these systems do not interfere with other applications, particularly if they are run at specific times or if batch runs are scheduled. When in doubt, add more paths (ISLs or trunks) through the fabric.**Clusters:** Clusters often have behavioral side effects that must be considered. This is particularly true during storage provisioning. It is possible, for example, for a cluster to inundate the fabric and storage arrays with LUN status queried and other short frame requests. This behavior can cause frame congestion in the fabric and can stress the control processors of the arrays. Make sure that you spread out the LUNs accessed by the hosts in the cluster across as many arrays as possible.
- **Congestion:** Traffic congestion (total link capacity regularly consumed) is remedied by adding more links or more members to a trunk. Frame congestion is typically addressed by dealing with the nodes causing the congestion.
- **Misbehaving devices:** It has been stated earlier that there is very little that a fabric can do to mitigate the effects of a badly behaving device, other than to remove it from the fabric. Brocade supports a Brocade FOS capability called Port Fencing, which is designed to isolate rogue devices from the network. Port Fencing works with Brocade Fabric Watch to disable a port when a specific threshold has been reached. Port Fencing, in combination with Bottleneck Detection, can be used for detecting and isolating high-latency devices from impacting the rest of the devices in the fabric.
- **Initiator and targets:** If possible, isolate host and storage ports on separate switches for much greater control over the types of controls that you can apply to misbehaving and high-latency devices. The effect on applications is typically much less severe if a host is disabled versus disabling a storage port, which may be servicing flows from many servers.

Monitoring

- **Brocade Network Advisor** is a powerful proactive monitoring and management tool that offers customizable health and performance dashboards to provide all critical information in a single screen. With Brocade Network Advisor, you can manage your entire network infrastructure.
- Use **Brocade Fabric Watch** to monitor switch and director resource consumption, port utilization, and port errors. Brocade Fabric Watch is also used to trigger Port Fencing.
- **Advanced Performance Monitoring** is an end-to-end monitoring tool that can help when you encounter congestion, including frame congestion.
- **Bottleneck Detection** is very useful in detecting latencies in devices and across links. It can help clarify whether high buffer credit zero counts are actually a problem. Once device latencies have been addressed, it is often useful to apply other controls, such as Port Fencing, to improve the resiliency of the fabric by isolating new misbehaving devices or future high latencies.
- **Brocade SAN Health** is a free utility that provides a lot of useful information to the storage or SAN administrator. You can look at ISL fan-in rations, get Visio diagrams of fabrics, verify firmware levels on switches, and find a host of other valuable information.

IOPS and VMs

Another method for determining bandwidth and/or oversubscription is to use the IOPS between host and storage devices. If the typical I/O size is known, along with the typical number of IOPS, then the administrator can calculate both average and estimated peak loads in terms of Megabytes per second (MB/sec). Next, look at the paths through the network for these I/Os, along with I/Os from other devices using the same network paths. Then use these data points to calculate bandwidth utilization and/or oversubscription ratios for devices and ISLs.

The use of VMs and the mobility of these VMs can make such IOPS calculations a challenge, as loads can shift when VMs move. Thus, the administrator needs to be aware of the potential VM loads on each physical server and their associated application loads for VMs.

While these calculations can certainly produce an accurate picture of bandwidth requirements for the storage network, they can be complicated even in a small network topology. This is why the simple approach discussed above is generally recommended.

ROUTED TOPOLOGIES—METASANS

The FC-FC routing service enables Fibre Channel SANs to share devices between two or more fabrics without merging those fabrics. The advantages for a routed topology are a reduced number of switch domains and zones for management within an edge fabric, fault isolation to a smaller edge fabric, interoperability with legacy fabrics, and increased security. In general, edge fabrics with Fibre Channel Routing (FCR) topologies follow the same basic best practice design guidelines as traditional fabrics, core-edge architectures for example. The FCR feature can be used for local or between fabrics across dark fiber or Wide-Area Networks (WANs) using FCIP.

Note: *Refer to Brocade Scalability Guidelines for FCR scalability limits.*

The primary considerations for using FCR are as follows:

- A limited number of LUNs shared between fabrics
- A limited number of servers that need to share LUNs between fabrics
- Share archiving devices like tape libraries
- The migration of legacy M-EOS or Brocade FOS fabrics to current Brocade FOS-based platforms
- OEM support
- Security separation in a managed services environment

There should be redundancy at the fabric, switch, and Inter-Fabric Link (IFL) levels (see Figures 17–19). A routed SAN, or MetaSAN, environment consists of multiple edge fabrics interconnected through one or more backbone fabrics. Multiple backbone fabrics are in parallel and belong to only the A or B fabric, not both. A core-edge topology can be used at both the backbone level and at the edge fabric level, such that edge fabrics and backbone fabrics are both deployed in a core-edge fashion.

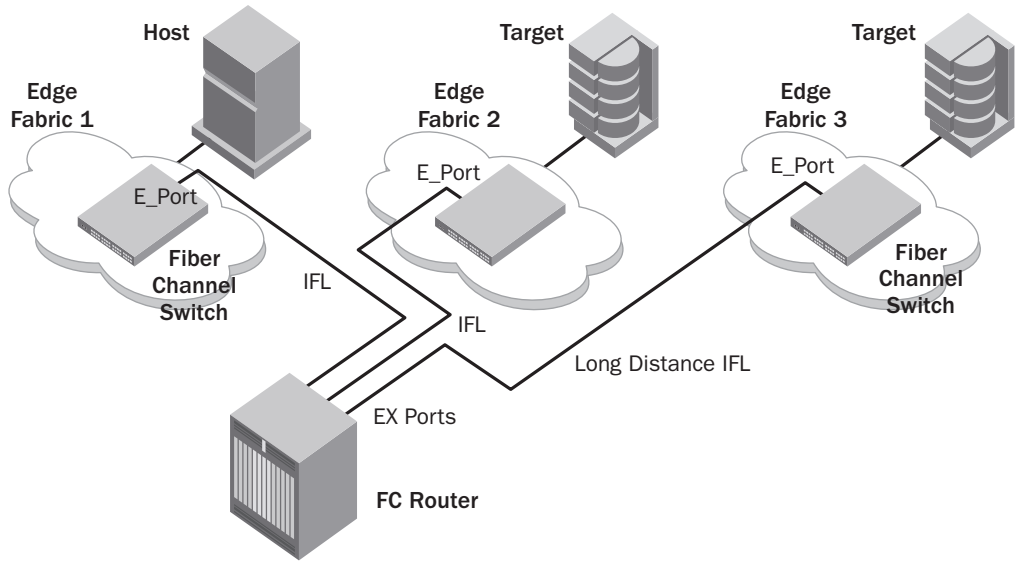

Figure 17. Typical MetaSAN topology.

The implementation and configuration of ISLs (and IFLs, in the case of FCR) should be based on the expected data flow between the switches and/or fabrics in question and the desired level of redundancy between edge switches and across the routed SAN. Below are some architectural examples of MetaSAN topologies.

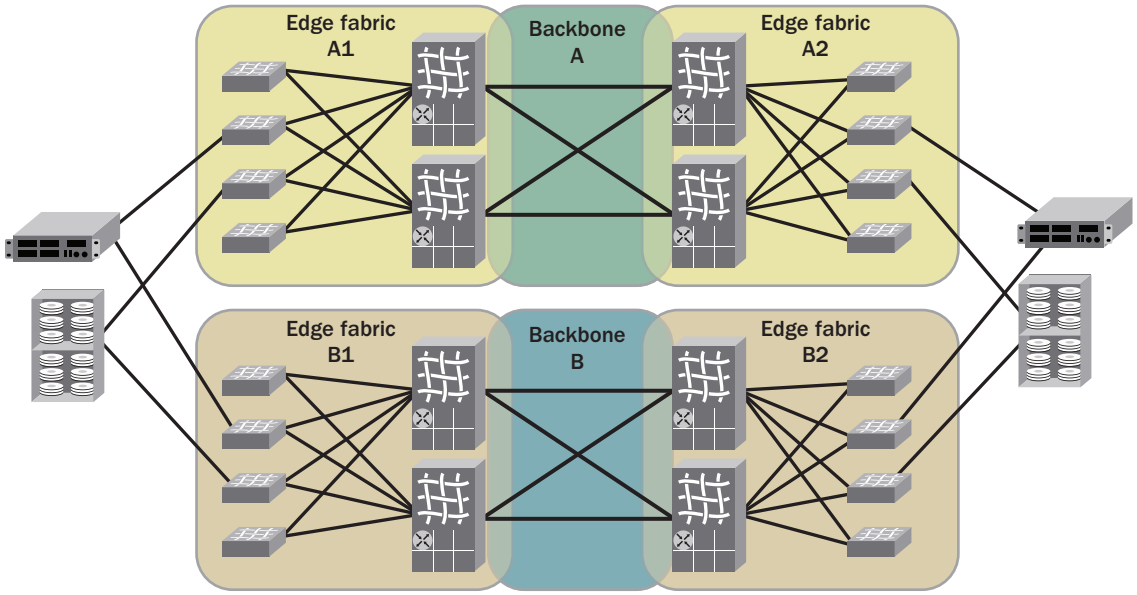

Figure 18. Example of a simple core-to-core attached backbone in a redundant routed fabric topology.

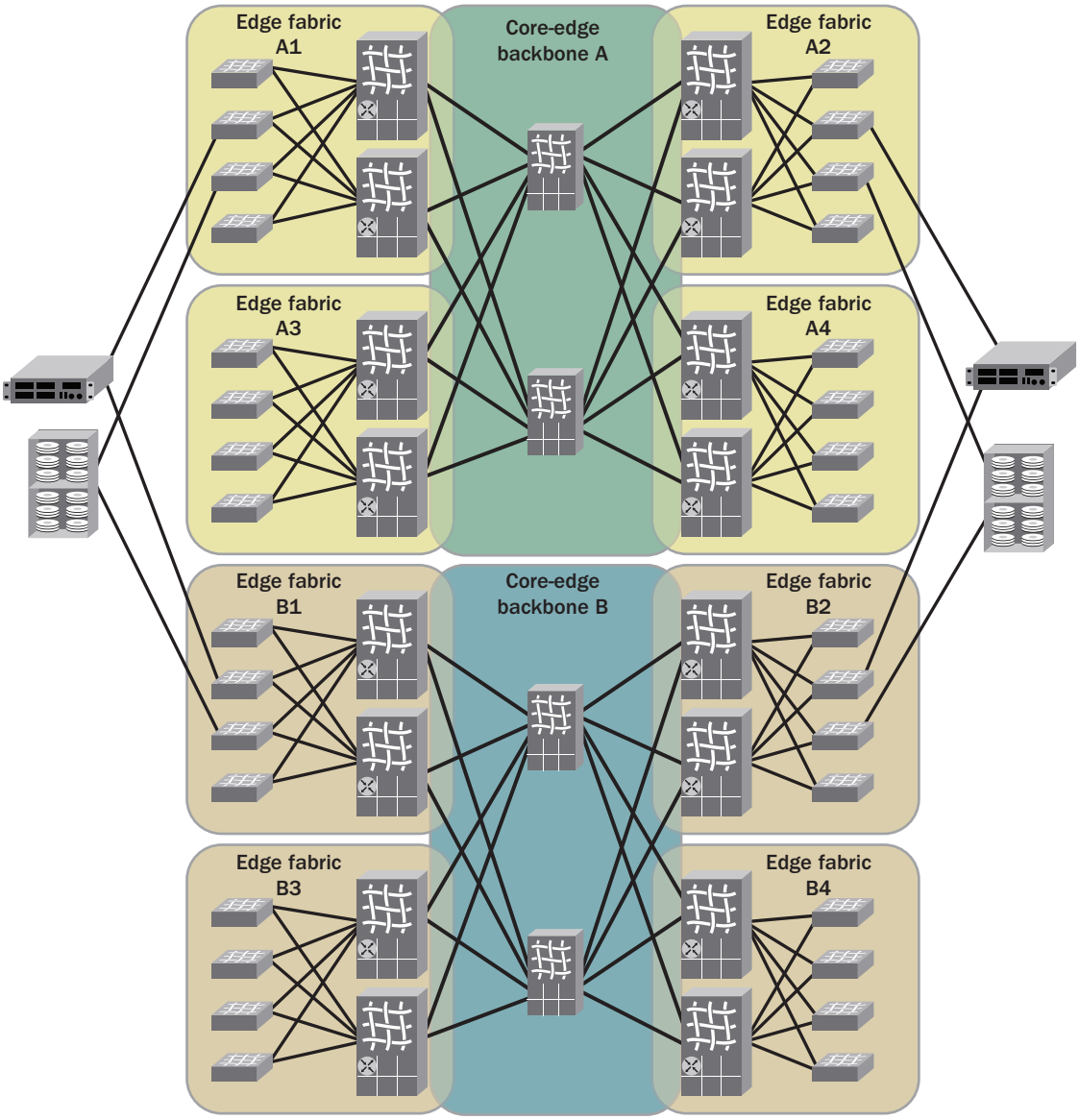

Figure 19. Example of an edge-core-edge backbone in a redundant routed fabric topology.

Backbone Considerations

There are many factors to consider when designing backbone fabrics. As mentioned above, the general SAN topology recommendations are applicable to backbone fabrics. There should be redundant fabrics, switches, and paths between the end-points (source and destination). Consider the following factors when identifying the best switch platforms and backbone topology, including switch interconnections:

- The number of edge fabrics impacts the backbone topology, as well as the manner in which edge fabrics are attached to the backbone. Brocade 8- and 16 Gbps platforms can support FCR functionality on all standard FC ports, and they provide a much more flexible solution when compared to legacy FCR platforms.
- **Composition of edge fabrics:**
  - **Legacy switches:** The presence of legacy Brocade switches anywhere in the SAN environment impacts the features that are supported and, depending on the platform and firmware version, may have other impacts as well.

  - **M-Series (legacy McDATA) switches:** Similar to legacy switches, the presence of M-Series switches anywhere in the SAN environment impacts the features that are supported and, depending on the platform and firmware version, may have other impacts as well.
  - **Advanced SAN application/features:** If you are considering implementing advanced SAN applications and/or features, the key factor is support (or compatibility) of the application on the SAN switch platforms being considered, as well as the ability to support those features across FCR.
- **Projected inter-fabric traffic patterns:**
  - **Quantity (bandwidth utilization):** You should provision a sufficient number of IFLs between each edge and the backbone to accommodate the projected traffic (at peak load) to and from each edge fabric. In addition, you should provision enough ISLs within the backbone to accommodate the projected traffic (at peak load) that will traverse the backbone.
  - **Bursty versus continuous traffic:** Bursty traffic is more forgiving than continuous traffic, since it generally handles temporary spikes in latency (unavailability of bandwidth). If the traffic pattern is largely made up of continuous streams of data, then provision extra bandwidth.

- **Small versus large frame size:** Fibre Channel is a high-speed, low-latency protocol. It relies, however, on buffer-to-buffer credits to handle flow control. This mechanism is a fundamental part of the Fibre Channel standard and ensures lossless connections. Thus, a series of 100 small frames uses the same number of buffers as a series of 100 large frames. Large frames, on the other hand, use more bandwidth. In other words, a large amount of small-frame traffic can fully utilize available buffers, while consuming only a very small amount of available bandwidth. Therefore, you need to consider not only bandwidth, but also the typical frame size. If the bulk of frames are expected to be smaller in size, then additional links and/or buffers should be allocated to the paths that will be handling those smaller frame I/O patterns. Pay extra attention to this type of congestion, because backbones can become congested and adversely impact the performance of all connected edge fabrics. When in doubt, overprovision IFLs.
- **Distance (location of fabrics):** The distance between the end-points of the data transmission is an issue of providing adequate buffers for the projected traffic, and all of the potential traffic flows that might traverse the long-distance link(s) need to be considered. Given that long-distance solutions generally already have increased latency (simple physics of time to cover distance), it is important that long-distance links be overprovisioned for capacity, such that unexpected spikes do not adversely impact the data flow or, potentially, the entire network.
- **Virtual Fabrics (VF):** If VF is enabled, the base switch is like a backbone switch, and a base fabric is like a backbone fabric. All switches in a backbone fabric must have the same backbone fabric ID, which must be unique from the edge fabric.

Note: *The VF Fabric ID is also the backbone, and Fabric ID and EX_Ports and VEX_Ports can be configured only on the base switch.*

- **Zoning:** Traffic Isolation (TI) zones and FCR: Some VE_Port-based features, such as tape pipelining, require the request and corresponding response traffic to traverse the same VE_Port tunnel across the MetaSAN. Use TI Zones to ensure that the request and response traverse the same VE_Port tunnel; you must set up TI Zones in the edge and backbone fabrics. In addition to setting up TI Zones, you must also ensure that the devices are in an LSAN zone so that they can communicate with each other. If failover is enabled and the TI path is not available, an alternate path is used. If failover is disabled and the TI path is not available, then devices are not imported.
- A potential for growth exists in the following:
  - **Number of fabrics:** If the number of fabrics is likely to increase, then deploy backbone fabrics such that they can readily accommodate additional edge fabrics and additional traffic loads.

- **Size of fabrics:** If the size of edge fabrics is likely to grow, and the inter-fabric traffic is expected to grow accordingly, provision additional IFLs and ISLs such that the capacity of available paths stays well ahead of current usage. That way, incremental growth on the edge can be accommodated without the need to immediately upgrade the backbone.
- **Amount of traffic between fabrics:** If the inter-fabric traffic is expected to grow even without growth in the individual edge fabrics, then provision additional IFLs and ISLs such that the capacity of available paths stays ahead of current usage. That way, incremental increases in data flow across the backbone can be accommodated without the need to immediately upgrade the backbone. Make sure that you allow for plenty of room for backbone expansion.

Avoiding Congestion

Just as with a flat Layer 2 fabric, a routed SAN needs to be evaluated for traffic bandwidth and potential bandwidth utilization between all end-points. For routed topologies, this means calculating traffic flowing in and out of every edge fabric and providing enough links into and across the backbone to accommodate that traffic. Use the same guidelines that apply to ISLs when connecting fabrics through IFLs for improved utilization and resiliency. As often happens as fabrics evolve, an edge fabric can be of higher performance versus the backbone, resulting in a completely oversubscribed backbone. This can lead to congestion at peak loads and high latency due to slow device response. Prior to upgrading the edge fabric, consider increasing the number of ISLs or upgrading the backbone to avoid congestion.

Available Paths

The best approach is to have multiple trunked paths between edge fabrics so that traffic can be spread across available resources; however, it is never good practice to attach both A and B fabrics to the same backbone router. From the perspective of FC, you should adhere to the concept of an “air gap” all the way from host to storage. A common device connected to both A and B fabrics can cause a SAN-wide outage. If an air gap is implemented, faults on one fabric cannot affect the other fabric. These faults can manifest from defects in host, fabric, or storage hardware and software, as well as human error. It is not relevant that FCR keeps the fabrics separate, because these types of faults can transcend FCR and cause the entire SAN to fail.

Design Guidelines and Constraints for Routed SANs

Some of the key metrics and rules of thumb for routed SAN topologies are:

- Keep A and B fabrics separated all the way from host to storage from a FC perspective. This is referred to as the “air gap.” This does not include IP networks passing FCIP traffic, although FCIP ISL end points should never cross-connect A and B fabrics, as this is the same as a traditional ISL cross connecting A and B fabrics.
- Localize traffic within an edge fabric as much as possible.
- Have a plan for predefining the domains in the fabric (for example, edge switches with a certain range, translate domains in a certain range that connect to the backbone fabric, and unique backbone fabric IDs to avoid domain overlap).
- Consider upgrading the backbone to higher performance prior to upgrading the edge fabric.
- When sharing devices using FCR between a core with B-Series switches running Brocade FOS and an edge fabric with M-Series switches running M-EOS, connect only one McDATA edge switch to the Brocade Backbone, or keep the number of EX_Ports from each FCR roughly balanced in edge fabric.
- No more than one long-distance hop between source and destination.
- Place long-distance links within the backbone (as opposed to between edge and backbone), as edge fabrics can then be isolated from disruption on the long-distance links. An edge fabric that contains a long-distance link is referred to as a remote edge. Remote edges can be the product of VEX_Ports and EX_Ports that connect to FC-based DWDM. Remote edges are not considered best practice.
- Use Logical SAN (LSAN) zones only for devices that will actually be communicating across the backbone. In other words, do not make every zone an LSAN zone for ease.

- As edge fabrics and the routed network grow, the use of "filters" such as LSAN zone binding and LSAN tagging can improve topology convergence time and efficient usage of FCR resources.
- Make the backbone fabrics redundant to improve resiliency. This means redundancy for each fabric; therefore, fabric A would be redundant and so would fabric B. Fabric B would never be used as the redundancy for fabric A, and vice-versa.
- Backbone fabrics that share connections to the same edge fabrics must have unique backbone fabric IDs. This statement is referring to the case in which there are multiple "A" fabrics and multiple "B" fabrics. This does not refer to sharing connections between A and B fabrics.
- TI Zones within the backbone fabric cannot contain more than one Destination Router Port (DRP) per each fabric.
- TI over FCR is supported only from edge fabric to edge fabric. Traffic isolation from backbone to edge is not supported.
- UltraScale ICLs on the core blade cannot be used for FCR.

VIRTUAL FABRICS TOPOLOGIES

The Brocade FOS Virtual Fabrics (VF) feature provides a mechanism for partitioning and sharing hardware resources, with the intention of providing more efficient use, deterministic paths for FCIP, increased fault isolation, and improved scalability. Virtual Fabrics use hardware-level fabric isolation between Logical Switches (LSs) and fabrics. Logical Fabrics consist of one or more Logical Switches across multiple physical switches (non-partitioned).

Hardware-level fabric isolation is accomplished through the concept of a Logical Switch, which provides the ability to partition physical switch ports into one or more "logical" switches. Logical Switches are then connected to form Logical Fabrics. As the number of available ports on a switch continues to grow, partitioning switches gives storage administrators the ability to take advantage of high-port-count switches by dividing physical switches into different Logical Switches. Without VF, an FC switch is limited to 256 ports. A storage administrator can then connect Logical Switches through various types of ISLs to create one or more Logical Fabrics.

There are three ways to connect Logical Switches: a traditional ISL, IFL (EX_Port used by FCR), and Extended ISL (XISL). An ISL can only be used for normal L2 traffic between the connected Logical Switches, carrying only data traffic within the Logical Fabric of which the ISL is a member. One advantage of Virtual Fabrics is that Logical Switches can share a common physical connection, and each LS does not require a dedicated ISL. In order for multiple Logical Switches, in multiple Logical Fabrics, to share an ISL, Virtual Fabrics supports an XISL connection, which is a physical connection between two base switches. Base switches are a special type of Logical Switch that are specifically intended for intra- and inter-fabric communication. As mentioned, base switches are connected via XISLs and form the base fabric.

Once a base fabric is formed, the Virtual Fabric determines all of the Logical Switches and Logical Fabrics that are physically associated via the base fabric, as well as the possible routes between them. For each local Logical Switch, a Logical ISL (LISL) is created for every destination Logical Switch in the same Virtual Fabric that is reachable via the base fabric. Thus, an XISL comprises the physical link between base switches and all of the virtual connections associated with that link. In addition to XISL support, the base fabric also supports IFLs via EX_Port connections for communication between Virtual Fabrics. Base switches also interoperate with FC router switches, either in the base fabric or in separate backbone fabrics.

VF Guidelines

If no local switching is used, any set of ports in the chassis/fabric can be used to create a Virtual Fabric. If local switching is used, ports for the VF fabric should be from the same port groups.

Use Case: FICON and Open Systems (Intermix)

Virtual Fabrics enable customers to share FICON and FCP traffic on the same physical platform. As chassis densities increase, this is a viable option for improved hardware utilization while maintaining director class availability. The primary reasons for moving to an Intermix environment are the following:

- Array-to-array RDR of FICON volumes (uses FCP)
- ESCON-FICON migration
- Sharing of infrastructure in a non-production environment
- Reduced TCO
- Growth of zLinux on the mainframe

From a SAN design perspective, the following guidelines are recommended when considering FICON Intermix:

- Connect devices across port blades (connectivity from the same device should be spread over multiple blades).
- One-hop count still applies (there are "Hops of No Concern" in some cases).

For details, see the Best Practices Guide: Brocade FICON/FCP Intermix.

INTELLIGENT SERVICES

In-Flight Encryption and Compression—Gen 5 Fibre Channel Platforms Only

Brocade Gen 5 Fibre Channel platforms support both in-flight compression and/or encryption at a port level for both local and long-distance ISL links. In-flight data compression is a useful tool for saving money when either bandwidth caps or bandwidth usage charges are in place for transferring data between fabrics. Similarly, in-flight encryption enables a further layer of security with no key management overhead when transferring data between local and long-distance data centers besides the initial setup.

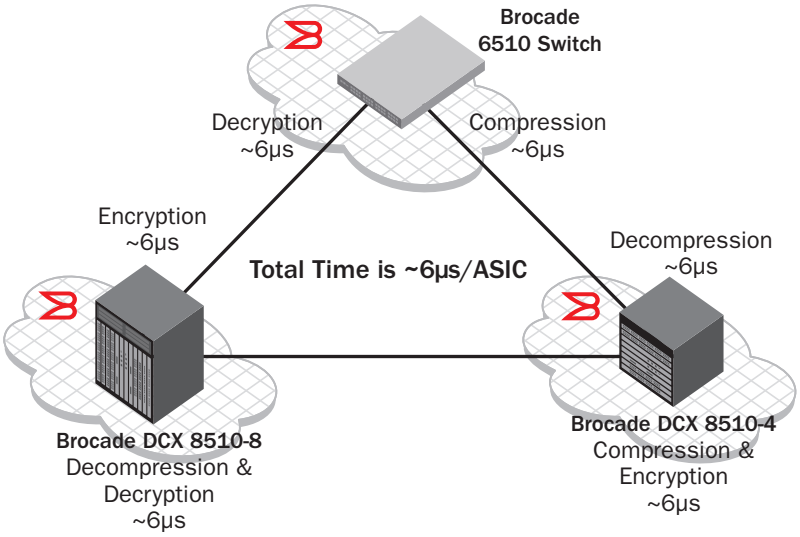

Figure 20: Latency for encryption and compression.

Enabling in-flight ISL data compression and/or encryption increases the latency as the ASIC processes the frame compression and/or encryption. Approximate latency at each stage (encryption and compression) is 6.2 microseconds. For example (see Figure 20), compressing and then encrypting a 2KB frame incurs approximately 6.2 microseconds of latency on the sending Condor3-based switch and incurs approximately 6.2 microseconds

of latency at the receiving Condor3-based switch in order to decrypt and uncompress the frame. This results in a total latency time of 12.4 microseconds, again not counting the link transit time.

Virtual Fabric Considerations (Encryption and Compression)
The E_Ports in the user-created Logical Switch, base switch, or default switch can support encryption and compression. Both encryption and compression are supported on XISL ports, but not on LISL ports. If encryption or compression is enabled and ports are being moved from one LS to another, it must be disabled prior to moving from one LS to another.

In-Flight Encryption and Compression Guidelines

- It is supported on E_Ports and EX_Ports.
- ISL ports must be set to Long-Distance (LD) mode when compression is used.
- Twice the number of buffers should be allocated if compression is enabled for long distance, as frame sizes may be half the size.
- If both compression and encryption are used, enable compression first.
- When implementing ISL encryption, using multiple ISLs between the same switch pair requires that all ISLs be configured for encryption—or none at all.
- No more than two ports on one ASIC can be configured with encryption, compression, or both when running at 16 Gbps speed. With Brocade FOS v7.1, additional ports can be utilized for data encryption, data compression, or both if running at lower than 16 Gbps speeds.
- Encryption is not compliant with Federal Information Processing Standards (FIPS).

DISTANCE EXTENSION TOPOLOGIES

For a complete DR solution, SANs are typically connected over metro or long-distance networks. In both cases, path latency is critical for mirroring and replication solutions. For native Fibre Channel links, the amount of time that a frame spends on the cable between two ports is negligible, since that aspect of the connection speed is limited only by the speed of light. The speed of light in optics amounts to approximately 5 microseconds per kilometer, which is negligible compared to typical disk latency of 5 to 10 milliseconds. The Brocade Extended Fabrics feature enables full-bandwidth performance across distances spanning up to hundreds of kilometers. It extends the distance ISLs can reach over an extended fiber by providing enough buffer credits on each side of the link to compensate for latency introduced by the extended distance.

Buffer Allocation

Buffer credits are a measure of frame counts and are not dependent on the data size (a 64 byte and a 2KB frame both consume a single buffer). Standard 8-Gb transceivers support up to 150 meters. (Refer to Appendix A for data rates and distances.) Users should consider the following parameters when allocating buffers for long-distance links connected via dark fiber or through a D/CWDM in a pass-thru mode:

1. Round-Trip Time (RTT)—in other words, the distance
2. Frame processing time
3. Frame transmission time

Some good general guidelines are:

- Number of credits = 6 + ((link speed Gb/s * Distance in KM) / frame size in KB)
- Example: 100 KM @2k frame size = 6 + ((8 Gb/s * 100) / 2) = 406
- Buffer model should be based on the average frame size
- If compression is used, number of buffer credits needed is 2x the number of credits without compression.

On the Brocade DCX 8510 with Gen 5 Fibre Channel platform, 4K buffers are available per ASIC to drive 16 Gbps line rate to 500 KM at 2KB frame size.

Brocade FOS v7.1 provides users additional control when configuring a port of an LD or LS link, allowing users to specify the buffers required or the average frame size for a long-distance port. Using the frame size option, the number of buffer credits required for a port is automatically calculated. These options give users additional flexibility to optimize performance on long-distance links.

In addition, Brocade FOS v7.1 provides users better insight into long-distance link traffic patterns by displaying the average buffer usage and average frame size via CLI. Brocade FOS v7.1 also provides a new CLI "portBufferCalc" that automatically calculates the number of buffers required per port given the distance, speed, and frame size. The number of buffers calculated by this command can be used when configuring the portCfgLongDistance command. If no options are specified, then the current port's configuration is considered to calculate the number of buffers required.

Note: *The ClearLink D_Port mode can also be used to measure the cable distance to a granularity of 5 meters between two 16 Gbps platforms; however, ports must be offline.*

Fabric Interconnectivity over Fibre Channel at Longer Distances

SANs spanning data centers in different physical locations can be connected via dark fiber connections using Extended Fabrics, a Brocade FOS optionally licensed feature, with wave division multiplexing, such as: Dense Wave Division Multiplexing (DWDM), Coarse Wave Division Multiplexing (CWDM), and Time Division Multiplexing (TDM). This is similar to connecting switches in the data center with one exception: additional buffers are allocated to E_Ports connecting over distance. The Extended Fabrics feature extends the distance the ISLs can reach over an extended fiber. This is accomplished by providing enough buffer credits on each side of the link to compensate for latency introduced by the extended distance. Use the buffer credit calculation above or the new CLI tools with Brocade FOS v7.1 to determine the number of buffers needed to support the required performance.

Any of the first 8 ports on the 16 Gbps port blade can be set to 10 Gbps FC for connecting to a 10 Gbps line card D/CWDM without the need for a specialty line card. If connecting to DWDMs in a pass-thru mode where the switch is providing all the buffering, a 16 Gbps line rate can be used for higher performance.

Recommendations include the following:

- Connect the cores of each fabric to the DWDM.
- If using trunks, use smaller and more trunks on separate port blades for redundancy and to provide more paths. Determine the optimal number of trunk groups between each set of linked switches, depending on traffic patterns and port availability.

FC over IP (FCIP)

Basic FCIP Architectures

Fibre Channel over IP (FCIP) links are most commonly used for Remote Data Replication (RDR) and remote tape applications, for the purpose of Business Continuance via Disaster Recovery. Transporting data over significant distances beyond the reach of a threatening event will preserve the data such that an organization can recover from that event. A device that transports FCIP is often called a channel extender.

RDR is typically storage array to array communications. The local array at the production site sends data to the other array at the backup site. This can be done via native FC, if the backup site is within a practical distance and there is DWDM or dark fiber between the sites. However, more commonly what is available is a cost-sensitive infrastructure for IP connectivity and not native FC connectivity. This works out well, because the current Brocade technology for FCIP is very high speed and adds only a minute amount (about 35 µs) of propagation delay, appropriate for not only asynchronous RDR and tape applications but also synchronous RDR applications. Best practice deployment of FCIP channel extenders in RDR applications is to connect the FC F_Ports on the channel extender directly to the FC N_Ports on the array, and not go through the production fabric

at all. On most large scale arrays, the FC port that has been assigned to RDR is dedicated to only RDR and no host traffic. Considering that the RDR port on the array can only communicate RDR traffic, there is no need to run that port into the production fabric. There are valid reasons to have to go through a production fabric such as IBM SVC, EMC CLARiiON (host shared FC port), tape applications, or connectivity of many arrays (more than can be accommodated by the Brocade 7800 Extension Switch alone) over the same WAN infrastructure. A single FCIP channel extender can be directly connected to both "A" and "B" controllers on the storage array, which is known as a two-box solution. Alternatively, a channel extender is dedicated to the "A" controller and a different channel extender dedicated to the "B" controller, which is known as a four-box solution. A single service provider for both the "A" and "B" paths can be used, or different service providers can be used, depending on the requirements of the organization and its tolerance to the recurring monthly costs of the additional WAN connection. Figure 21 shows the two-box solution, and Figure 22 shows the four-box solution with a single service provider.

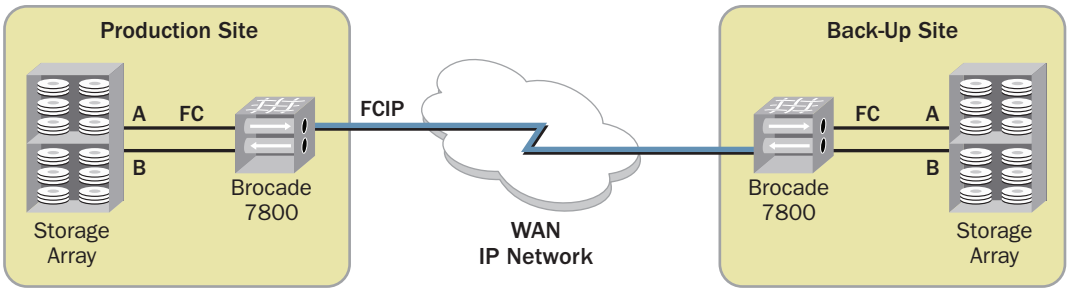

Figure 21. Two-box RDR solution directly connected to the storage array.

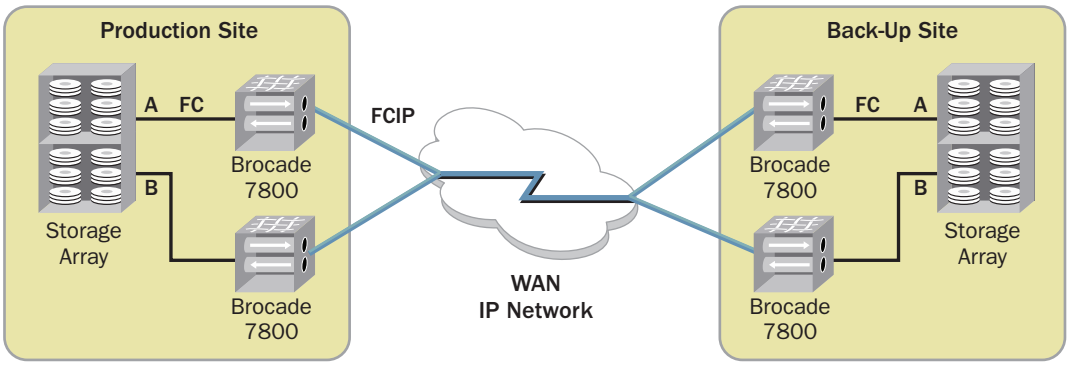

Figure 22. Four-box RDR solution directly connected to the storage array.

It is also possible to incorporate the production fabrics into the FCIP path, although this is not recommended unless there is a compelling reason to do so (Figure 23). Tape is most often the reason for attaching channel extenders to a production fabric, providing greater connectivity to many devices. IBM SVC has requirements for connectivity to the production fabric. Many customers prefer to connect channel extenders to the production fabric when using EMC CLARiiON because MirrorView and SAN Copy applications that run on the array share bandwidth with hosts over the same FC port; therefore, the production fabric is used to direct the traffic to the hosts and to the channel extender.

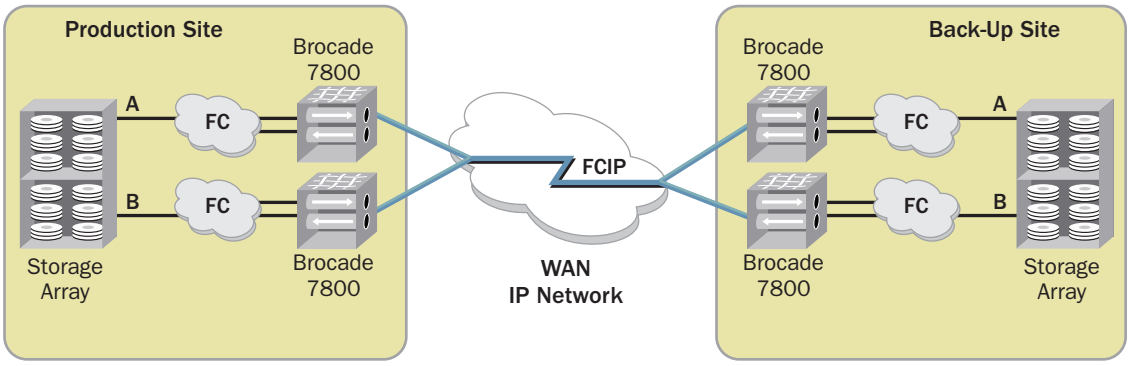

Figure 23. Four-box solution connected to production fabrics.

In environments that require production fabric attached channel extenders, it is not best practice to connect the same channel extender to both "A" and "B" fabrics. The best practice is to have two redundant FC fabrics in all production environments in which an organization would suffer losses if the SAN were to go down. Even a momentary outage can "blue screen" or hang servers, causing them to have to be rebooted, which can take a significant amount of time in some situations. The division of the "A" and "B" fabrics implies that there is an air gap between the two autonomous fabrics all the way from the server to the storage array. There are no physical data links between the two independent fabrics. The servers are equipped with FC software drivers (for example, Microsoft MPIO, EMC PowerPath) for their HBAs that monitor the individual paths sending data across all of them. If a path is detected as down, the driver will fail over the traffic to the remaining path(s). This is best practice for maximum availability. This implies that a single channel extender that must connect via the production fabric cannot connect to both the "A" and "B" fabrics simultaneously, as shown in Figure 24. If no Fibre Channel Routing (FCR) is being used, the fabric would merge into one big fabric, which clearly destroys any notion of an A and B fabric. If FCR is used, the fabrics do not merge; however, there is still a device with a common Linux kernel attached to both fabrics. This is not acceptable if maximum availability is the goal, and it is considered a poor practice with high risk. Brocade does not recommend this type of architecture. This type of architecture, having a common device connected to both the A and B fabrics, is also susceptible to human error, which can also bring down the entire SAN (meaning both A and B fabrics).

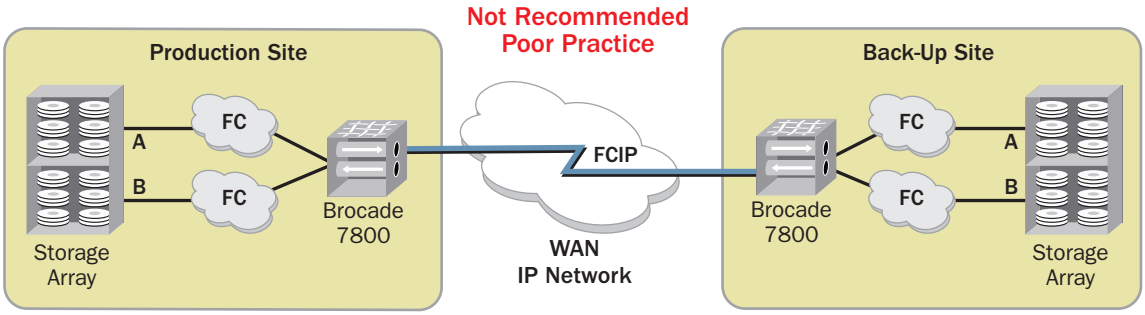

Figure 24. Poor practice two-box solution connected to production fabrics.

When connecting channel extenders to production fabrics, each production fabric should be designed using best practice concepts in a traditional core-edge fashion, with the core tier including either the connections to standalone channel extenders such as the Brocade 7800 or the FCIP-capable blades, such as the Brocade FX8-24 Extension Blade. Each channel extender should be connected to a fabric using at least two parallel FC ISLs, as shown in Figure 23.

When using a four-box solution, it is inappropriate to make ISL cross-connections between the two channel extenders within a data center site and both the “A” and “B” FC fabrics, because of the reasons discussed above. However, it is permissible to do so on the Ethernet/WAN side (see Figure 25).

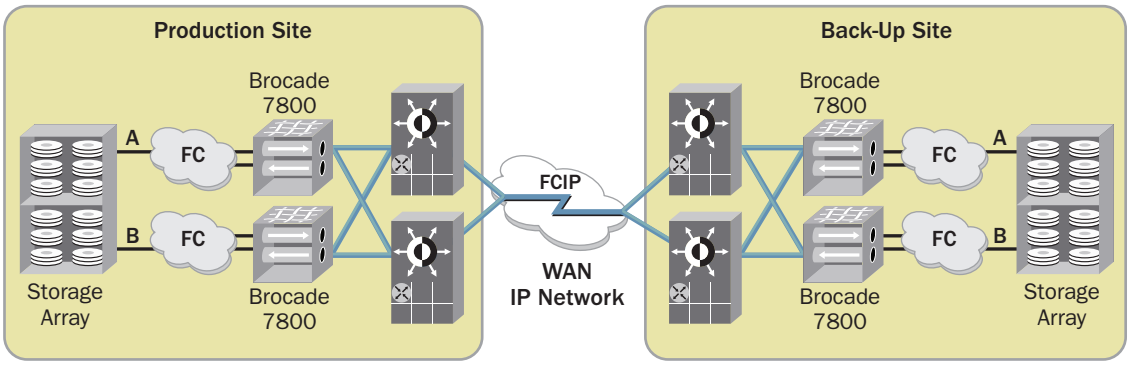

Figure 25. Ethernet connectivity to dual-core WAN infrastructure.

FCIP with FCR

The FCIP tunnel traditionally traverses a WAN or IP cloud, which can have characteristics that adversely impact a Fibre Channel network. The FCIP link across a WAN is essentially an FC ISL over an IP link. In any design, it should be considered an FC ISL. Repeated flapping of a WAN connection can cause disruption in directly connected fabrics. This disruption many come about from the many fabric services trying to reconverge, and reconverge again, and reconverge again, over and over. This causes the CPU on the switch or director to max out. If the CPU can no longer process the various tasks required to operate a fabric, there may be an outage. If you limit the fabric services to within the local fabric itself and do not allow them to span across the WAN, you can prevent this from occurring. FCR provides a termination point for fabric services, referred to as a “demarcation point.” EX_Ports and VEX_Ports are demarcation points in which fabric services are terminated, forming the “edge” to the fabric. A fabric isolated in such a way is referred to as an “edge fabric.” There is a special case in which the edge fabric includes the WAN link because a VEX_Port was used; this type of edge fabric is referred to as a “remote edge fabric.”

FCR does not need to be used unless there is a production fabric that must be isolated from WAN outages. When connecting to array ports directly for RDR, FCR provides no benefit. Mainframe environments are precluded from using FCR, as it is not supported with FICON.

Please note, when a mainframe host writes to a volume on the Direct Access Storage Device (DASD), and that DASD performs RDR to another DASD, then DASD to DASD traffic is not using FICON. It is an open systems RDR application such as EMC SRDF, HDS HUR or TrueCopy, IBM Metro Mirror or Global Mirror, or HP Continuous Access. These open-system RDR applications can use FCR, even though the volumes they are replicating were written by the FICON host.

There are some basic FCR architectures:

- First and simplest, no FCR or one big fabric: this type of architecture is used with the mainframe and when the channel extenders are directly connected to the storage arrays.
- Second, edge-backbone-edge, in which edge fabrics bookend a transit backbone between them.
- Third, when a VEX_Port is used, the resulting architecture can be either backbone-remote edge or edge-backbone-remote edge, depending on whether devices are connected directly to the backbone or an edge fabric hangs off of the backbone. Both are possible.

Using EX_Ports and VEX_Ports

If an FCR architecture is indicated, an "X" port is needed. An "X" port is a generic reference for an EX_Port or a VEX_Port. The only difference between an EX_Port and a VEX_Port is that the "V" indicates that it is FCIP-facing. The same holds true for E_Ports and VE_Ports; VE_Ports are E_Ports that are FCIP-facing.

The best practice in an FC routed environment is to build an edge fabric to backbone to edge fabric (EBE) topology. This provides isolation of fabric services in both edge fabrics. This architecture requires an EX_Port from the Backbone to connect to an E_Port in the edge fabric, as shown in Figure 26. The backbone fabric will continue to be exposed to faults in the WAN connection(s), but because its scope is limited by the VE_Ports in each edge fabric, and since edge fabric services are not exposed to the backbone, it does not pose any risk of disruption to the edge fabrics in terms of overrunning the CPUs or causing a fabric service to become unavailable. The edge fabric services do not span the backbone.

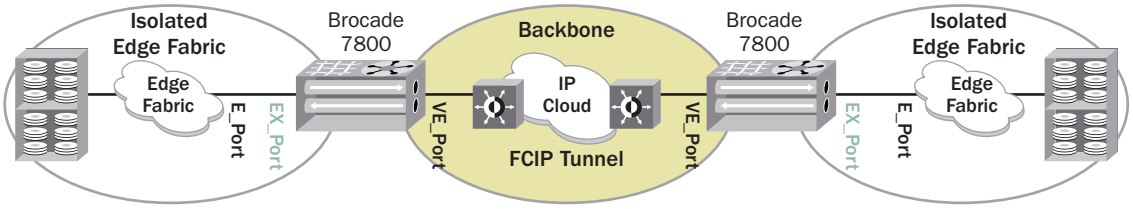

Figure 26. Edge-backbone-edge FCR architecture.

There may be cases in which an EBE topology cannot be accommodated; alternatively, the main production fabric can be isolated from aberrant WAN behavior, while allowing the backup site to remain exposed. This provides a greater degree of availability and less risk compared to not using FCR at all. This type of architecture uses VEX_Ports that connect to a remote edge fabric. The important point to observe here is that the remote edge fabric continues to be connected to the WAN, and the fabric services span the WAN all the way to the EX_Port demarcation point. This means that the fabric services spanning the WAN are subject to disruption and repeated reconvergence, which can result in an outage within the remote edge fabric. This may not be of great concern if the remote edge fabric is not being used for production (but merely for backup), since such WAN fluctuations are not generally ongoing.

There are two topologies that you can build from remote edge fabrics. In the first, as shown in Figure 27, production devices are attached directly to the backbone. In the second, as shown in Figure 28, the backbone connects to a local edge fabric. In both cases, the other side is connected to a remote edge fabric via a VEX_ Port. Also in both cases, the production fabrics are isolated from the WAN. Between the two architectures, the second architecture with the edge fabric is recommended for higher scalability. The scalability of connecting devices directly to the backbone is relatively limited.

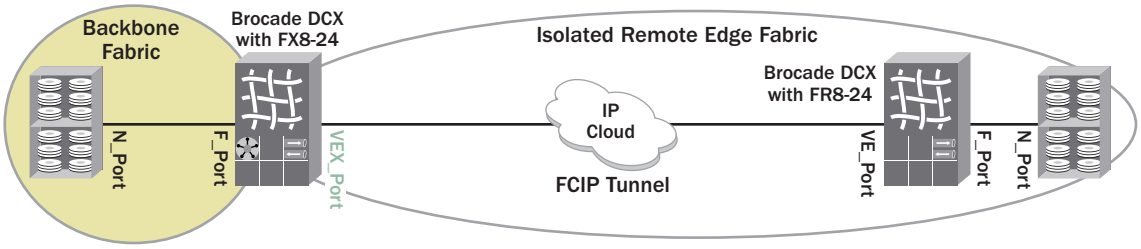

Figure 27. Backbone-remote edge architecture.

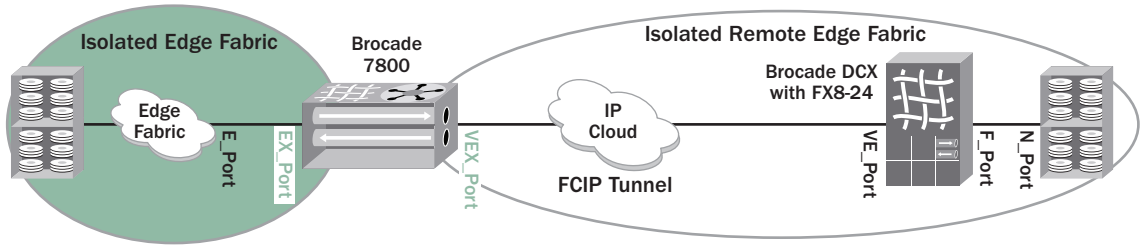

Figure 28. Edge-remote edge architecture.

Another design consideration with "X" ports is: How many can be in a path? This is indeed limiting. If you start from inside an FC Router (refer to Figure 29) and move toward the initiator or target, you may only pass through 1 "X" port along the way. If you pass through 2 "X" ports to get to the initiator or target, the architecture is not supported.

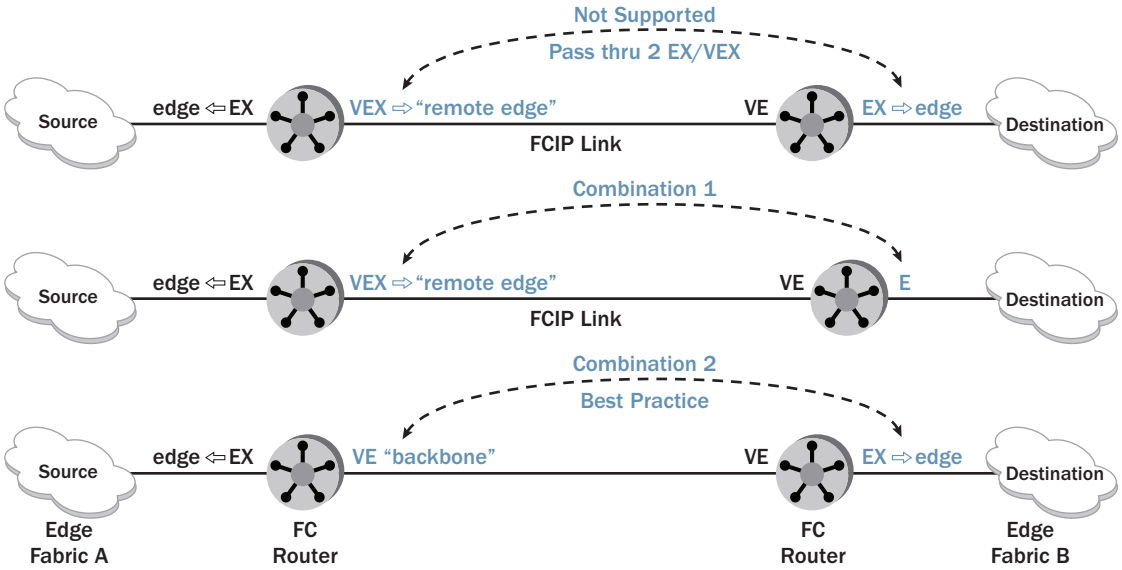

Figure 29. "X" ports along a path.

The Integrated Routing (IR) license, which enables FCR on Brocade switches and directors, is needed only on the switches or directors that implement the "X" ports. Any switches or directors that connect to "X" ports and have no "X" ports of their own do not need the IR license. The IR license is not needed on the E_Port/VE_Port side to connect to the EX_Port/VEX_Port side.

FCIP with FICON Environments

In a mainframe environment, FCR cannot be used, because the FICON protocol is not supported. Most often, VFs are used on a Brocade DCX platform forming multiple LSs. One LS is designated as the fabric to be extended. The other LSs are used for production traffic. Only channels and control units that must communicate across the channel extenders are connected to this LS. These LS fabrics merge across the WAN, forming a single fabric. This is acceptable because the number of devices is limited to just those that are communicating across the WAN, which tend not to be in production. In other words, they are backup tape or an RDR application—namely IBM XRC. If the production fabrics were to be exposed to the WAN they might be vulnerable to the type of outage previously described.

IBM does not consider a connection to a channel extender to be a hop of concern. Only one hop is supported by IBM, and a connection to a channel extender is not considered a hop. Brocade UltraScale ICLs are also not considered hops. The key here is—what is considered a channel extender? A channel extender is a Brocade 7800 with no end devices connected to any of the FC ports. Only ISLs are allowed to be connected to the FC ports. If a FICON director were to connect to a Brocade DCX with a Brocade FX8-24 blade in it, and that blade was in its own LS with no other end devices connected to the LS, that would be considered a valid architecture and no hop would be counted, even if end devices were connected to other LSs within the same Brocade DCX.

Figure 30 shows a typical mainframe extension architecture.

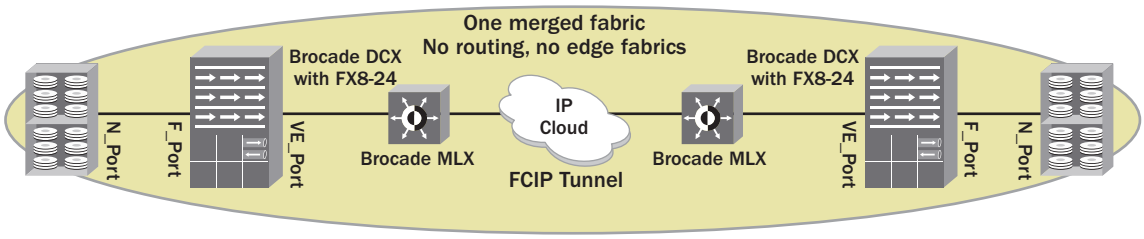

Figure 30. Typical mainframe extension architecture.

Advanced FCIP Configuration

Beyond the physical topology layout, there are many additional features and functions associated with FCIP connections. These include: IP Security (IPsec), compression, Adaptive Rate Limiting (ARL), and more. There are definite advantages to using these features. See the SAN extension product documentation for details.

IPsec

With the Brocade 7800/FX8-24, it is always prudent to enable IPsec. All data leaving a data center and going into an infrastructure that guarantees no security (no service provider will guarantee your data) should be encrypted to prevent man-in-the-middle attacks. The Brocade design goals of IPsec were to make it as practical to deploy as it is in WiFi. Would your company operate WiFi with no encryption? No, of course not. IPsec operates at line rate and is HW-based. There are no additional licenses or costs to use IPsec on Brocade. It adds an insignificant amount of latency at 5 µs. The setup is easy. Configuration is easy by establishing a Pre-Shared Key (PSK) on both sides. Brocade IPsec uses all the latest encryption technologies such as: AES 256, SHA-512 HMAC, IKEv2, and Diffie-Hellman. The key is regenerated approximately every 2 GB of data that passes across the link, and that process is not disruptive.

Compression

Compression is recommended in every type of architecture, including those built for RDR/S. There are three modes of compression besides off:

Mode 1, Brocade optimized Lempel–Ziv (LZ), is a hardware-implemented compression algorithm that is suitable for synchronous applications because it adds a mere 10 µs of added latency. In addition, Brocade LZ can accommodate the maximum ingress rate for which the Brocade 7800/FX8-24 has been built, so it is line rate and poses no bottleneck for ingress traffic. LZ typically gets about a 2:1 compression ratio.

Mode 2, Brocade optimized Dynamic Huffman Coding, is a software with hardware assist compression algorithm. Software-based algorithms are not suitable for synchronous applications, because they add too much processing latency. Brocade Dynamic Huffman Coding can accommodate up to 8 Gbps ingress from the FC side. For the Brocade 7800, that means 8 Gbps for the entire box. For the Brocade FX8-24 blade, that means 8 Gbps for each FCIP complex, of which there are two, one for each 10 GbE interface. The 10 GbE interfaces belong to the complex for 10 GbE interface 1 (XGE1). Mode 2 has been designed to work efficiently with an OC-48 WAN connection. Mode 2 typically gets about a 2.5:1 compression ratio.

Mode 3, Deflate, also known as GZIP, is entirely a software-based algorithm and not suitable for synchronous applications. Deflate takes the tradeoff between compression ratio and compression rate further. The maximum rate per FCIP complex is 2.5 Gbps ingress from the FC side. Mode 3 has been designed to work efficiently with an OC-12 WAN connection. Mode 3 typically gets about a 4:1 compression ratio.

Brocade makes no guarantees or promises as to the actual compression ratio your specific data will achieve. Many customers have achieved the typical values listed here.

Adaptive Rate Limiting (ARL)

ARL is a technology that should be an integral part of an FCIP network design whenever there is more than one FCIP interface feeding into the same WAN connection, or when the WAN is shared with other traffic. These are the most common use cases.

Each circuit is configured with a floor and ceiling bandwidth (BW) value. The bandwidth for the circuit will never be less than the floor value and never be more than the ceiling value. The bandwidth available to the circuit can be automatically adjusted between the floor and ceiling, based on conditions in the IP network. A congestion event causes the rate limit to adjust down towards the floor. An absence of congestion events causes it to rise up to the ceiling. ARL adjustments do not take place rapidly, which prevents massive congestion events from occurring. If the bandwidth is somewhere in the middle, ARL will make periodic attempts to adjust upward, but if it cannot, because of a detected congestion event, it will remain stable.

When more than one FCIP interface is feeding a WAN link, the two FCIP flows equalize and utilize the total available bandwidth. If one of the interfaces or boxes goes offline, such as when the interface is on a separate box, then ARL can readjust to utilize the bandwidth that is no longer being used by the offline interface. This maintains good utilization of the WAN bandwidth during periods of maintenance and box or optics failures.

In Figure 31, the blue circuit is feeding the WAN, after which the yellow circuit comes online. The blue and yellow circuits find equilibrium, as their aggregate bandwidth is equal to the available WAN bandwidth. When the yellow circuit goes offline again, the bandwidth is freed up and the blue circuit intermittently tests for that bandwidth and increase the rate limiting to take advantage of it. This continues until the ceiling is reached again.

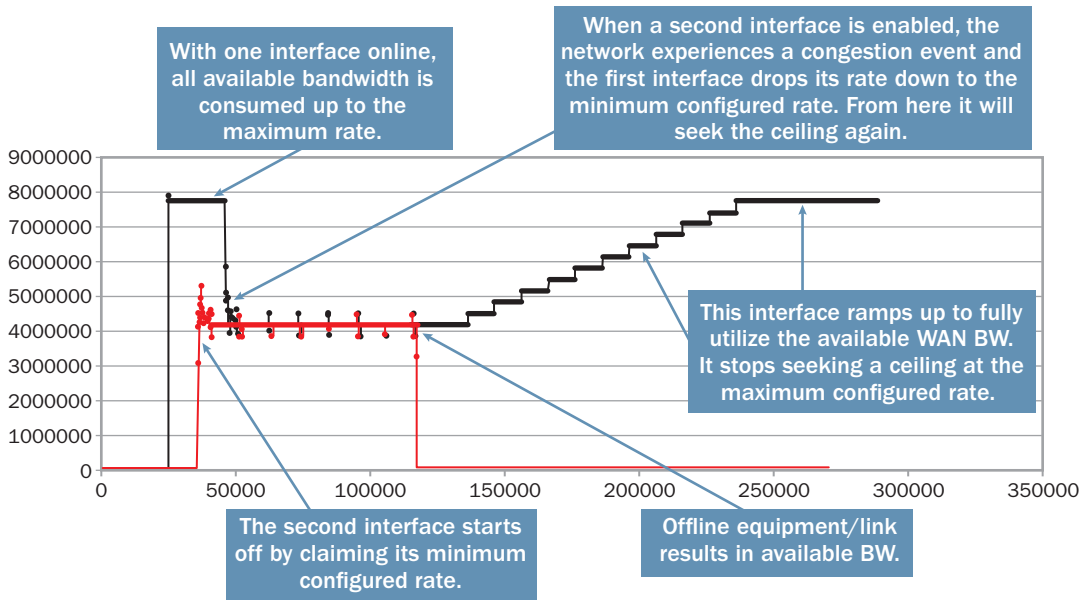
With one interface online, all available bandwidth is consumed up to the maximum rate.
When a second interface is enabled, the network experiences a congestion event and the first interface drops its rate down to the minimum configured rate. From here it will seek the ceiling again.
This interface ramps up to fully utilize the available WAN BW. It stops seeking a ceiling at the maximum configured rate.
The second interface starts off by claiming its minimum configured rate.
Offline equipment/link results in available BW.
9000000
8000000
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
0
50000
100000
150000
200000
250000
300000
350000

Figure 31. Adaptive Rate Limiting behavior for two flows.

In a shared link situation, if you think of the bandwidth as separated into three areas, black (0 → x bps), gray (x → y bps), and white (y → maximum bps), ARL can help manage the bandwidth usage. Black is the floor value for ARL. This is the amount of bandwidth that is reserved exclusively for FCIP. White is the ceiling value, and it is reserved exclusively for other shared traffic. Gray is the area in between, which FCIP may use if other shared traffic is not using it. This other shared traffic can also be another FCIP application such as tape. Black would be the RDR traffic; white would be tape traffic, and they would adaptively share the gray area. There are many ways in which you can use ARL. These are just a few popular examples.

PerPriority-TCP-QoS

PP-TCP-QoS is a Brocade innovation (patent pending).

Differentiated Services Code Point (DSCP) is an IP based (L3) Quality of Service (QoS) marking; therefore, since IP is end-to-end protocol, DSCP is an end-to-end QoS marking. DSCP has 64 values; however, the range of values from 0 through 63 do not denote the lowest priority through the highest priority. The valuing system works differently. First, all odd numbers are available for private use and can be used in any way that enterprise deems valuable. These odd numbers are for private use the same way that RFC 1918 IP addresses are; for example, 192.168.0.1 and 10.1.2.3 are private IP addresses that can be used in any way an enterprise wishes.

For non-private DSCP values, DSCP value 46 is referred to as Expedited Forwarding and is the highest priority. Zero is the default, and it is the lowest priority. There are four groups of High/Medium/Low (H/M/L) values referred to as Assured Forwarding. Another group of numbers has backwards compatibility with legacy ToS (Type of Service). The selection of DSCP to be used in the IP network is the responsibility of the IP network administrators. Without their buy-in and configuration of the Per-Hop Behavior (PHB) associated with QoS, no QoS can actually happen. The default behavior of Ethernet switches is to replace ingress QoS values with the default value (0), unless the data coming in on that interface is explicitly deemed to be QoS "trusted." This prevents end users from setting their own QoS values unannounced to the IP networking administrators.

802.1P is a data link-based (L2) QoS marking; therefore, the scope extends only from the interface of one device to the interface of the directly attached device. Devices that enforce 802.1P provide QoS across that data link. 802.1P has a header that resides in the 802.1Q VLAN tagging header; therefore, VLAN tagging is required to get 802.1P QoS marking. Brocade FOS refers to 802.1P as L2CoS. There are only eight values for 802.1P—from 0 to 7. Zero is the lowest priority and the default. Seven is the highest priority.

The Brocade 7800/FX8-24 supports three levels of priority (H/M/L). The default amount of BW that the scheduler apportions during times of contention is 50/30/20 percent. QoS portioning of BW occurs only during times of contention; otherwise, the BW is shared equally across all priorities. It is possible to change the default portions to any values you wish, as long as High>Middle>Low, and the aggregate of all the priorities equals 100 percent.

There are four TCP sessions per FCIP circuit: H, M, L, and F-Class. F-Class uses a strict queuing, which means that if there is any F-Class traffic to send, it all gets sent first. There is very little F-Class traffic, and it does not interfere with data traffic. Each TCP session is autonomous and does not rely on other TCP sessions or settings. Each TCP session can be configured with its own DSCP, VLAN tagging, and 802.1P values. This permits that TCP session (priority) to be treated independently in the IP network from site-to-site, based on the SLA for that QoS priority.

Brocade has QoS in Brocade FC/FICON fabrics and across FC ISLs via Virtual Channels (VCs). There are different VCs for H/M/L/F-Class, each with its own set of Buffer-to-Buffer Credits and flow control. There are five VCs for high levels, four VCs for medium levels, and 2 VCs for low levels. Devices are assigned to QoS VCs by enabling QoS on the fabric and then putting the letters QOSH_ or QOSL_ as a prefix to the zone name. The default is QOSM_, so there is no need to explicitly designate medium zones. Once devices are assigned to these VCs, they use these VCs throughout the fabric. If data ingresses to a Brocade 7800/FX8-24 via an ISL on a particular VC, the data is automatically assigned to the associated TCP sessions for that priority. Devices that are directly connected to the Brocade 7800/FX8-24 are also assigned to the associated TCP session priority based on the zone name prefix.

DSCP and L2CoS are configured on a per-FCIP circuit basis. It is recommended that you not alter the QoS markings for F-Class traffic unless it is required to differentiate and expedite F-Class traffic across the IP network between the sites. Failure of F-class traffic to arrive in a timely manner will cause instability in the FC fabric. This is less of an issue with directly connected separate RDR networks. FCIP networks that have to be connected to the production FC fabrics can use FCR (IR license) to protect the edge fabrics from instability.

FCIP Design Best Practices

Bandwidth Allocation

For RDR, best practice is to use a separate and dedicated IP connection between the production data center and the backup site. Often a dedicated IP connection between data centers is not practical. In this case, bandwidth must at least be logically dedicated. There are a few ways this can be done. First, use QoS, and give FCIP a high priority. This logically dedicates enough bandwidth to FCIP over other traffic. Second, use Committed Access Rate (CAR) to identify and rate-limit certain traffic types. Use CAR on the non-FCIP traffic to apportion and limit that traffic to a maximum amount of bandwidth, leaving the remainder of the bandwidth to FCIP. Set the aggregate FCIP rate limit on the Brocade 7800 switch or FX8-24 blade to use the remaining portion of the bandwidth. This results in logically dedicating bandwidth to FCIP. Last, it is possible, with massive overprovisioning of bandwidth, for various traffic types to coexist over the same IP link. Brocade FCIP uses an aggressive TCP stack called Storage Optimized TCP (SO-TCP), which dominates other TCP flows within the IP link, causing them to back off dramatically. If the other flows are UDP-based the result is considerable congestion and excessive dropped packets for all traffic.

Best practice is to always rate limit the FCIP traffic on the Brocade 7800 or FX8-24 blade and never rate limit FCIP traffic in the IP network, which often leads to problems that are difficult to troubleshoot. The rate limiting technology on the Brocade 7800/FX8-24 is advanced, accurate, and consistent, so there is no need to double rate limit. If policy required you to double rate limit, then the IP network should set its rate limiting above that of the Brocade 7800/FX8-24 with plenty of headroom.

To determine the amount of network bandwidth needed, it is recommended that a month's worth of data is gathered using various tools that are host-, fabric-, and storage-based. It is important to understand the host–to-disk traffic, since that is the amount of traffic to be replicated, or mirrored, to the remote disk.

If you are going to be doing synchronous RDR (RDR/S), then record peak values. If you are going to be using asynchronous RDR (RDR/A) then record the average value over the hour. RDR/S must have enough bandwidth to send the write I/O immediately; therefore, there must be enough bandwidth to accommodate the entire demand, which is peak value. RDR/A needs only enough bandwidth to accommodate the high average discovered over an adequate recording period, because RDR/A essentially performs traffic shaping, moving the peaks into the troughs, which works out to the average. It cannot be the average over a very long period of time, because those troughs may not occur soon enough to relieve the array of the peaks. This causes excessive journaling of data, which is difficult to recover from.

Plot the values into a histogram. More than likely, you will get a Gaussian curve (see Figure 32). Most of the averages will fall within the first standard deviation of the curve, which is 68.2% of the obtained values. The second standard deviation will include 95.4% of the obtained values, which are enough samples to determine the bandwidth you will need. Outside of this, the values are corner cases, which most likely can be accommodated by the FCIP network due to their infrequency. Use a bandwidth utilization value that you are comfortable with between σ and 2σ. You can plan for a certain amount of compression, such as 2:1. However, best practice is to use compression as a way to address future bandwidth needs. It is probably best not to push the limit right at the start, because then you will have nowhere to go in the near future.

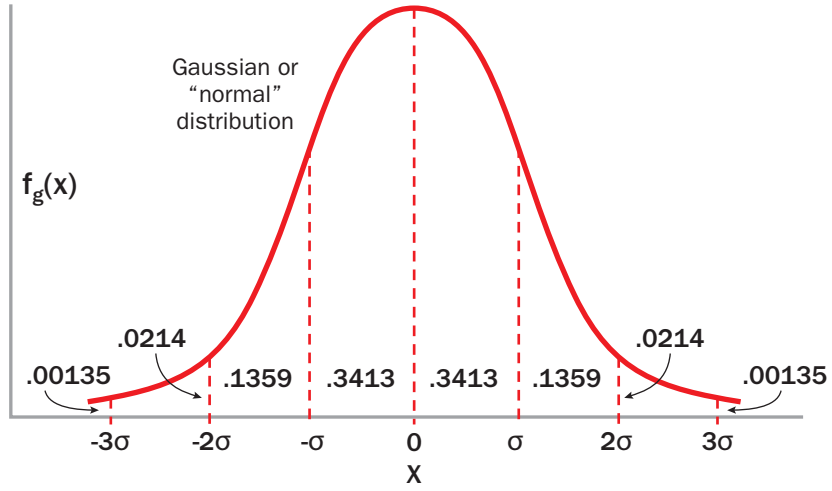

Figure 32. Gaussian curve.

For remote tape/Virtual Tape Library (VTL) backup, the extension systems can be used to extend the data center fabric from the location where backup servers are attached to the remote fabric and where the remote tape/VTL is located, as shown by the blue arrow in Figure 33.

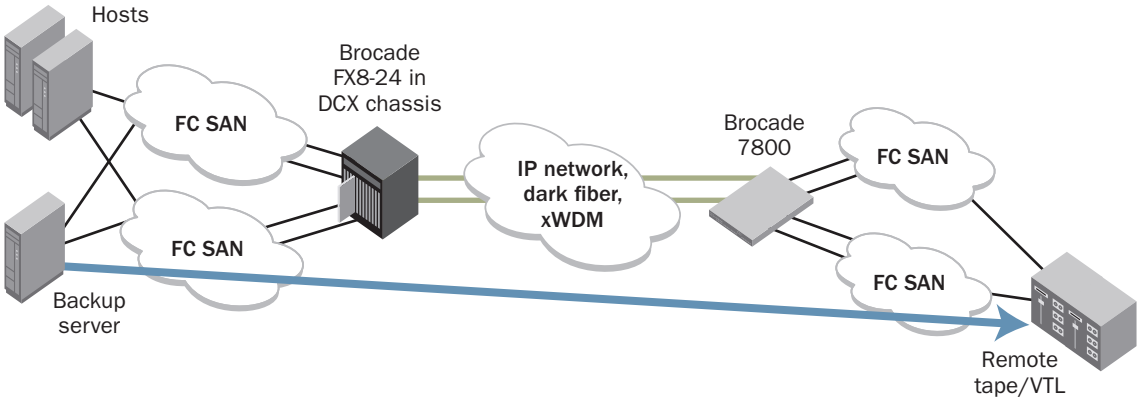

Figure 33. SAN extension extending the data center fabric to remote tape or VTL.

It is important that you measure the tape volume in MB/h, as well as the number of tape drives used, and to determine the batch window in hours. These measurements will determine the amount of network bandwidth that is required.

You can take advantage of FCIP Trunking to implement redundant network routes from site to site. But it is important to understand whether traffic can fail over to the alternate route transparently or whether that will impact traffic flow.

For both disk and tape extension using emulation (FastWrite for disk and Tape Pipelining for tape), a single tunnel between sites is recommended. If multiple tunnels must be used, use Traffic Isolation (TI) zones or logical switch configuration to ensure that the same exchange always traverses by the same tunnel in both directions. Use multiple circuits instead of multiple tunnels for redundancy and failover protection.

FCIP Trunking

The Brocade 7800 and FX8-24 have an exclusive feature called FCIP Trunking. FCIP Trunking offers the ability to perform the following functions:

- Bandwidth aggregation
- Lossless failover/failback
- Granular load balancing
- In-order delivery
- Prevention of IFCC on mainframes

A single tunnel defined by a VE_Port or VEX_Port may have one or more circuits associated with it. A circuit is an FCIP connection defined by a source and destination IP address and other arguments that define its characteristics, such as compression, IPsec, QoS, rate limit, VLAN tag, and others. All the circuits terminate at the single VE/VEX_Port on each side; therefore, there are no multiple tunnels or ISLs, but only a single tunnel load balanced across multiple circuits. The one ISL that an FCIP Trunk forms is from VE_Port to VE_Port or VEX_Port to VE_Port.

The circuits can have different characteristics. They can have different RTTs and take different paths and different service providers. They can have different bandwidths up to 4x. This means that if one circuit is an OC-3, the most the other circuit(s) can be is OC-12, because the bandwidth delta is 4x.

FCIP Trunking is considered best practice in most cases. For example, consider the architecture shown in Figure 34. The FC perspective has already been discussed in detail. Here, consider the Ethernet/IP perspective and how FCIP Trunking pertains to a high availability design.

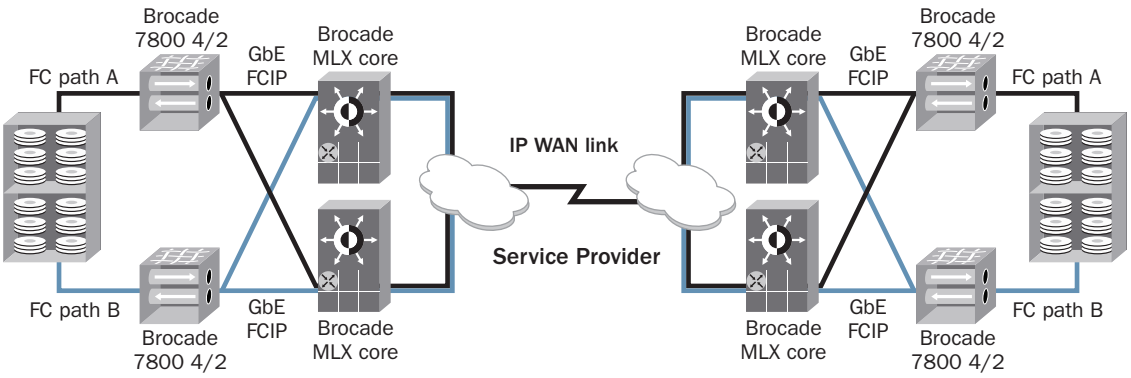

Figure 34. Four Brocade 7800 High Availability Architecture.

Virtually all data centers have redundant IP core routers/switches. It is best practice to connect each Brocade 7800/FX8-24 to each of the IP core routers/switches for redundancy and resiliency purposes, as shown in Figure 34. Without FCIP Trunking this design would require two VE_Ports per Brocade 7800 4/2. There are two VE_Ports available in a 4/2; however, from a performance, resiliency, and redundancy point of view, this is not the best solution. Instead, it is better to use a single VE_Port on the 4/2 with FCIP Trunking. The VE_Port forms an FCIP tunnel with the opposing VE_Port, and there are two member circuits. Any FCIP tunnel with more than one circuit is called an FCIP Trunk. FCIP circuits are assigned to Ethernet interfaces and, in this case, each circuit is assigned to its own dedicated Ethernet interface. The Ethernet interfaces are then physically connected to an Ethernet switch/IP core router. One of the Ethernet interfaces is connected to core A, and one is connected to core B. Now there are two circuits that will load balance across both data center cores. With FCIP Trunking, if any of the following occurs the result is no loss of data: core routers fail or have to go offline for maintenance, bad Ethernet SFP or optical cable, and sub-second failover within the WAN network.

ARL is used to manage the bandwidth going into the cores based on the available WAN bandwidth. There may be a single WAN connection or separate WAN connections between the sites. ARL is used to manage the BW from the Brocade 7800s to the WAN connection. This example has a single WAN connection, although you could just as well use more than one WAN connection. ARL is configured such that the floor value is set to the WAN BW ÷ the number of interfaces feeding the WAN; in this case, it is 4 (2 from each Brocade 7800). The ceiling value is set to either the line rate of the GE interface or the available WAN BW. For example, if the WAN is an OC-12 (622 Mbps), the ceiling ARL value is set to 622 Mbps. The floor value is set to 155 Mbps. When all the interfaces are up and running, they will run at 155 Mbps. In an extreme case in which three Ethernet interfaces are offline, the remaining FCIP Ethernet interface will run at 622 Mbps, continuing to utilize all the WAN BW and keeping the RDR application satisfied.

All circuits have a metric of 0 or 1 associated with them, as shown in Figure 35. 0 is the preferred metric and is used until all metric 0 circuits have gone offline. After all circuits with metric 0 have gone offline, then metric 1 circuits are used. This is most useful with ring topologies, in which one span of the ring is used with metric 0 circuits and, if the span fails, then the other span is used with metric 1 circuits. Both metric 0 and 1 circuits can belong to the same FCIP Trunk (same VE_Port), which means that if the last metric 0 circuit fails and a metric 1 circuit takes over, no data in-flight is lost during the failover using LLL.

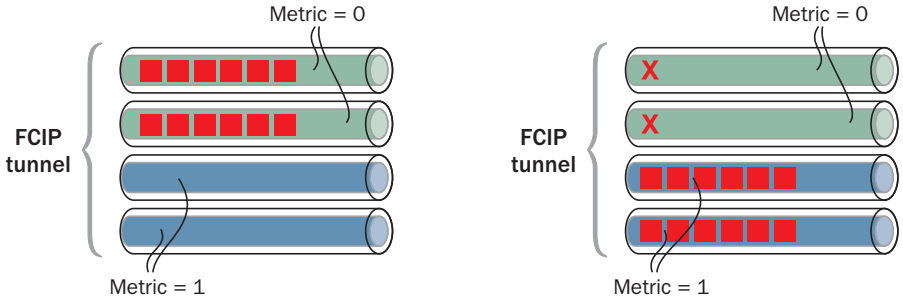

Figure 35. FCIP trunk circuits with metrics.

Brocade FCIP uses keepalives to determine circuit health. Keepalives are sent at the timer value divided by 5. Each keepalive that arrives resets the count. If the counter reaches 5, the circuit is deemed offline and goes down. Massive IP network congestion and dropped packets can conceivably cause all five keepalives to be lost in transit, causing the circuit to go down. You do not want the keepalive timer to be set too short, because the TCP sessions across the WAN have the ability to ride through very short outages and recover quickly. If the timer is too short, this will not happen before going down, although a longer keepalive interval will take longer to detect a bad circuit. FICON and FCP circuits have different default keepalive timer settings when they are configured. An argument that indicates FICON must be added when you are configuring circuits that are used for FICON. FICON has stricter timing than FCP and must have no more than a 1-second keepalive timer. FCP has more flexibility, and the default is 10 seconds; nevertheless, best practice is to also set the keepalive timer to 1 second unless the IP network tends to have congestion and deep buffers that inadvertently trigger FCIP circuit drops.

Protocol Optimization

Brocade FCIP offers seven different protocol optimizations: FICON tape read/write, FICON XRC, FICON Teradata, OSTP read/write, and FastWrite. With design practices, the most important concept to understand with protocol optimization is that there must be a determinant path for both outbound and return communications of an exchange. This means that all FC frames from an exchange must pass through the same two VE_Ports in both directions. This is difficult to achieve if there is a fabric on each end and multiple VE_Ports across which those fabrics can choose to send the data. In this case, there is no deterministic path.

Protocol optimization requires keeping the state of the protocol local to each end device. These states are kept in what is called a state machine. Brocade extension equipment can maintain state for up to about 20,000

state machines at any one time. A state machine lives for the life of an exchange, and then it is deleted. FICON Emulation operates somewhat differently, and it is not discussed here in detail. State machines live in the VE_Port and require the outbound and return of an exchange's sequences to pass through the same VE_Port; otherwise, the state of the protocol errors, and a failure occurs. A sequence that passes through one set of VE_Ports on the way out sets up state on those specific VE_Ports. If the sequence returns on a different set of VE_Ports, there is no state, and the VE_Ports cannot determine what that sequence is and which exchange it belongs to. To prevent this problem from occurring, the outbound and return path should be deterministic. There are a couple of methods for creating deterministic paths for protocol optimization.

First and most common, provide only one physical path. If a single VE_Port connects fabrics on the end points of an FCIP ISL, then there is only one physical path, and protocol optimization will work well. With FCIP Trunking, one VE_Port does not mean one FCIP connection, as there may be multiple FCIP circuits and Ethernet ports used to connect up the VE_Ports. This provides the resiliency and redundancy.

Second, if more than one VE_Port and protocol optimization are required, then you can use VF Logical Switches to provide a deterministic path. By dividing the Brocade 7800 or DCX Backbone into separate LSs and assigning a single VE_Port to that LS, a deterministic path across the Logical Fabric can be formed. Using VE_Ports on an FX blade can provide 10 Gbps of FCIP to one LS. Of course, the downside to this method is that there is no ubiquitous connectivity across all ports on the Brocade 7800 or DCX Backbone. The FC/FICON/VE ports become isolated, which is why there is a deterministic path.

Third, you can use TI zones to confine data to deterministic paths specific to ports from end-to-end.

Virtual Fabrics

The Brocade 8510 Backbone with the FX8-24 Extension Blade and the Brocade 7800 Extension Switch (4/2 and 16/6 models) all support VFs with no additional license. The Brocade 7800 supports a maximum of four Logical Switches and does not support a Base Switch. Because there is no Base Switch, the Brocade 7800 cannot provide support for XISL or FCR (no EX_Ports and VEX_Ports). FICON CUP is supported by IBM on two LSs only. VF on the Brocade 7800 must be disabled if a separate RDR network is not feasible and FCR is required to connect to production edge fabrics.

VF on the Brocade 7800/FX8-24 plays a primary role in providing ways to achieve deterministic paths for protocol optimization, or for the purposes of specific configuration and management requirements providing unique environments for FICON and FCP. Virtual Fabrics is the preferred alternative over TI Zones to establish deterministic paths necessary for protocol optimization (FCIP-FW, OSTP, and FICON Emulation). Protocol optimization requires that an exchange and all its sequences and frames pass through the same VE_Port for both outbound and return. This means that only a single VE_Port should exist within a VF LS. By putting a single VE_Port in an LS there is only one physical path between the two LS that are connected via FCIP. A single physical path provides a deterministic path. When a large number of devices or ports are connected for transmission across FCIP, as would be the case with tape for example, it is difficult to configure and maintain TI Zones, whereas it is operationally simplistic and more stable to use VF LS.

Configuring more than one VE_Port, one manually set with a higher FSPF cost, is referred to as a "lay in wait" VE_Port and it is not supported for FCIP-FW, OSTP or FICON Emulation. A "lay in wait" VE_Port can be used without protocol optimization and with RDR applications that can tolerate the topology change and some frame loss. A small number of FC frames may be lost when using "lay in wait" VE_Ports. If there are multiple VE_Ports within an LS, routing across those VE_Ports is performed according to the APTpolicy.

Virtual Fabrics are significant in mixed mainframe and open system environments. Mainframe and Open System environments are configured differently and only VFs can provide autonomous LSs accommodating the different configurations. Keep in mind that RDR between storage arrays is open systems (EMC SRDF, HDS Universal Replicator/TrueCopy, IBM Metro/Global Mirror, and HP Continuous Access), even when the volume is written by FICON from the mainframe.

Here is a list of configuration differences between FICON and Open System environments that require VF LS when mixing environments on the same switch or Director:

- The APTpolicy setting for FICON is not the same as Open Systems. FICON typically uses Port-Based Routing (PBR) or Domain-Based Routing (DBR) with Lossless enabled. Open Systems uses Exchange-Based Routing (EBR) without Lossless enabled.
- In-Order Delivery (IOD) is used in FICON environments. IOD is disabled in Open Systems environments.
- Security ACLs are required in cascaded FICON environments. Security ACLs are not used in Open Systems environments.
- FICON Management Server (FMS) with Control Unit Port (CUP) is enabled only on FICON LSs.
- Brocade Network Advisor management of the LS in FICON mode vs. Open Systems mode.

Understand that using a VE_Port in a selected LS does not preclude that VE_Port from sharing an Ethernet interface with other VE_Ports in other LSs. This is referred to as Ethernet Interface Sharing, refer to the next section.

Ethernet Interface Sharing

An FCIP Trunk uses multiple Ethernet interfaces by assigning the circuits that belong to that trunk to different Ethernet interfaces. Ipif are configured with IP addresses, subnet masks and an Ethernet interface, which assigns the ipif to the interface. When the FCIP circuit is configured, the source IP address has to be one that was used to configure an ipif, which in turn assigns the FCIP circuit to that Ethernet interface. It is possible to assign multiple IP addresses and circuits to the same Ethernet interface by assigning multiple ipif to that same interface, each with its own unique IP address.

Any one circuit cannot be shared across more than one Ethernet interface. An IP address/ipif/circuit can belong only to one Ethernet interface. Thus, if more than one Ethernet interface is desired, you must use multiple circuits. If the same IP address is attempted to be configured on more than one ipif, an error will occur, rejecting the configuration.

It is possible to share an Ethernet interface with multiple circuits that belong to different VF LSs. The Ethernet interface must be owned by the default switch (context 128). The ipif and iproute must also be configured within the default switch. The VE_Port is assigned to the LS you want to extend with FCIP and is configured within that LS. The FCIP tunnel is also configured within that LS using the IP addresses of the ipif that are in the default switch. This permits efficient use of the 10 GbE interfaces.

Often, for purposes of redundancy and resiliency, an FCIP Trunk has circuits that extend out of both of the 10 GbE interfaces. Each 10 GbE interface (XGE) has “native” VE_Ports from one of the two groups (xge1:12-21 or xge0:22-31). If you wish to extend a circuit from VE_Port 12 through xge0, you must use something called a cross-port. A cross-port requires an ipif and iproute that have been configured and explicitly designated for cross-port use; otherwise, the circuit cannot be configured for the non-native 10 GbE interface. By merely designating the ipif and iproutes to be used with non-native XGE interfaces, you can configure this type of circuit.

Figure 36 shows an example of two VF LS (50 and 60) on a Brocade FICON Director. The Ethernet interfaces are in the default switch LS context 128. There are two circuits per FCIP Trunk. There is a red FCIP Trunk and a blue FCIP Trunk. VE_Port 12 has one circuit that goes to its native xge1 interface and one circuit that goes to cross-port xge0. VE_Port 22 has one circuit that goes to its native xge0 interface and one circuit that goes to cross-port xge1. There are two 5 Gbps circuits emanating from each 10 GbE interface and transmitted over a physical Ethernet link to an IP core gateway router within the data center.

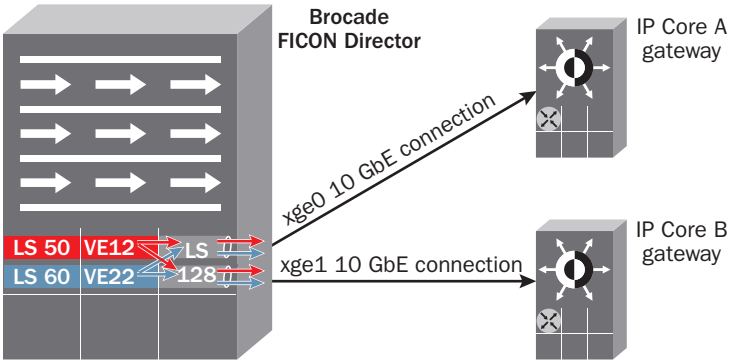

Figure 36. Two Virtual Fabric Logical Switches deployed on a Brocade FICON Director

Figure 37 shows in more detail the anatomy of the circuits from the VE_Ports in each LS. The iproute that is configured has the destination subnet/mask and associates that destination with a local gateway on the IP core router. In this example, each circuit requires its own iproute:

- Red IP core router A GW
- Red IP core router B GW cross-port
- Blue IP core router A GW cross-port
- Blue IP core router B GW

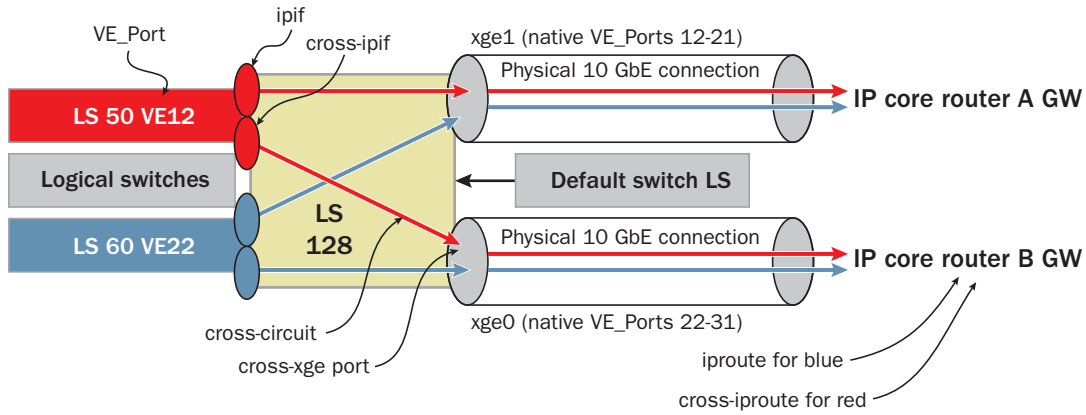

Figure 37. Detailed view of Logical Switch VE_Ports

What is Not Supported?

There are a few features that are not supported by Brocade:

- Cisco E_Port (ISL) interoperability is not supported by Brocade, and Brocade does not support any OEM that supports Cisco interoperability.
- VE_Ports or VEX_Ports that "lay in wait" are not supported for FICON, tape, and protocol optimizations. The concept here is that by manually adjusting the FSPF cost, you can provide a backup VE_Port upon an active VE_Port going offline. This may possibly be the case if there are two FX blades and one is taken offline. The problem is that there is considerable state information involved with FICON, tape, and protocol optimization processes that cannot be recovered from by merely bringing up a new VE_Port. These processes have to be restarted.

- Per-Packet Load Balancing (PPLB) is not supported by Brocade. PPLB is often the cause of chronic and sometimes severe Out-Of-Order Packets (OOOP). To some degree, OOOP can easily be handled by TCP; however, PPLB will cause TCP to go into overdrive and put a dramatic strain on the system. PPLB also causes large latencies in delivering sequences to the upper layer protocols, because TCP is waiting for packets that are arriving out of order. PPLB may also cause a significant number of retransmits that will unnecessarily use up valuable bandwidth. Moreover, there is no redeeming added value that PPLB brings to the network, as there are other load balancing techniques that do not cause such problems and that equally load the WAN links. It is for these reasons that Brocade does not support PPLB for FCIP WAN connections.

WORKLOADS

Many different kinds of traffic traverse a SAN fabric. The mix of traffic is typically based on the workload on the servers and the effect that behavior has on the fabric and the connected storage. Examples of different types of workload include these:

- **I/O-intensive, transaction-based applications:** These systems typically do high volumes of short block I/O and do not consume a lot of network bandwidth. These applications usually have very high-performance service levels to ensure low response times. Care must be taken to ensure that there are a sufficient number of paths between the storage and hosts to ensure that other traffic does not interfere with the performance of the applications. These applications are also very sensitive to latencies.
- **I/O-intensive applications:** These applications tend to do a lot of long block or sequential I/O and typically generate much higher traffic levels than transaction-based applications (data mining). Depending on the type of storage, these applications can consume bandwidth and generate latencies in both storage and hosts that can negatively impact the performance of other applications sharing their storage.
- **Host High Availability (HA) clustering:** These clusters often treat storage very differently from standalone systems. They may, for example, continuously check their connected storage for data integrity reasons and put a strain on both the fabric and the storage arrays to which they are attached. This can result in frame congestion in the fabric and can cause performance problems in storage arrays.
- **Host-based replication:** Host-based replication causes traffic levels to increase significantly across a fabric and can put considerable pressure on ISLs. Replicating to poorer-performing storage (such as tier 1-to-tier 2 storage) can cause application performance issues that are difficult to identify. Latencies in the slower storage can also cause "back pressure," which can extend back into the fabric and slow down other applications that use the same ISLs.
- **Array-based replication:** Data can be replicated between storage arrays as well.

Workload Virtualization

The past three years have witnessed a huge growth in virtualized workload. Available on IBM mainframes for decades, workload virtualization was initially popularized on Intel-based platforms by VMware ESX Server (now vSphere). Windows, UNIX, and Linux server virtualization is now ubiquitous in enterprise infrastructures. Microsoft now has a viable product with Hyper-V.

Most recently, organizations have started adopting workload virtualization for desktops. This technology is still in development but is evolving rapidly. (Desktop virtualization storage access is not addressed in this document.)

Intel-Based Virtualization Storage Access

Intel-based VMs typically access storage in two separate ways:

- They use some sort of distributed file system that is typically controlled by the hypervisor (the control program that manages VMs). This method puts the onus on the hypervisor to manage the integrity of VM data. All VM I/O passes through an I/O abstraction layer in the hypervisor, which adds extra overhead to every I/O that a VM issues. The advantage to this approach is that many VMs can share the same LUN (storage), making storage provisioning and management a relatively easy task. Today the vast majority of VMware deployments use this approach, deploying a file system called Shared VMFS.

- They create separate LUNs for each data store and allow VMs to access data directly through N_Port ID Virtualization (NPIV). The advantage of this approach is that VMs can access data more or less directly through a virtual HBA. The disadvantage is that there are many more LUNs to provision and manage.

Most VMs today tend to do very little I/O—typically no more than a few MB/sec per VM via very few IOPS. This allows many VMs to be placed on a single hypervisor platform without regard to the amount of I/O that they generate. Storage access is not a significant factor when considering converting a physical server to a virtual one. More important factors are typically memory usage and IP network usage.

The main storage-related issue when deploying virtualized PC applications is VM migration. If VMs share a LUN, and a VM is migrated from one hypervisor to another, the integrity of the LUN must be maintained. That means that both hypervisors must serialize access to the same LUN. Normally this is done through mechanisms such as SCSI reservations. The more the VMs migrate, the potentially larger the serialization problem becomes. SCSI reservations can contribute to frame congestion and generally slow down VMs that are accessing the same LUN from several different hypervisor platforms.

Design Guidelines

- If possible, try to deploy VMs to minimize VM migrations if you are using shared LUNs.
- Use individual LUNs for any I/O-intensive applications such as SQL Server, Oracle databases, and Microsoft Exchange.

Monitoring

- Use Advanced Performance Monitoring and Brocade Fabric Watch to alert you to excessive levels of SCSI reservations. These notifications can save you a lot of time by identifying VMs and hypervisors that are vying for access to the same LUN.

Unix Virtualization

Virtualized Unix environments differ from virtualized Windows deployments in a few significant ways.

First, the Unix VMs and hypervisor platforms tend to be more carefully architected than equivalent Windows environments, because more mission-critical applications have traditionally run on Unix. Frequently the performance and resource capacity requirement of the applications are well understood, because of their history of running on discrete platforms. Historical performance and capacity data will likely be available from the Unix performance management systems, allowing application architects and administrators to size the hypervisor platforms for organic growth and headroom for peak processing periods.

Second, VM mobility is not common for workload management in Unix deployments. VMs are moved for maintenance or recovery reasons only. IBM clearly states, for example, that moving VMs is limited to maintenance only. Carefully architected hypervisor/application deployments contain a mix of I/O-intensive, memory-intensive, and processor-intensive workloads. Moving these workloads around disturbs that balance and potentially leads to performance problems. Problem determination also becomes more difficult once VM migrations have to be tracked.

Third, virtualized mission-critical Unix applications such as large SQL Server database engines typically do much more block I/O than their Windows counterparts, both in volume and in transaction rates. Each hypervisor platform now produces the aggregate I/O of all those mission-critical applications. Backups, especially if they are host-based through backup clients, are also a serious architectural concern.

Recent Changes

Two technical advances create profound changes to storage deployments for mission-critical Unix applications: NPIV and storage virtualization.

Consider the IBM AIX VIO platform as an example to explain Unix workload virtualization. (Other vendor systems such as Oracle/Sun Solaris and HP HP-UX behave somewhat differently.) NPIV came late to Unix, with IBM recently adopting NPIV in AIX VIO 2.1 to improve traffic through the SCSI I/O abstraction layer. The difference is illustrated in Figure 38.

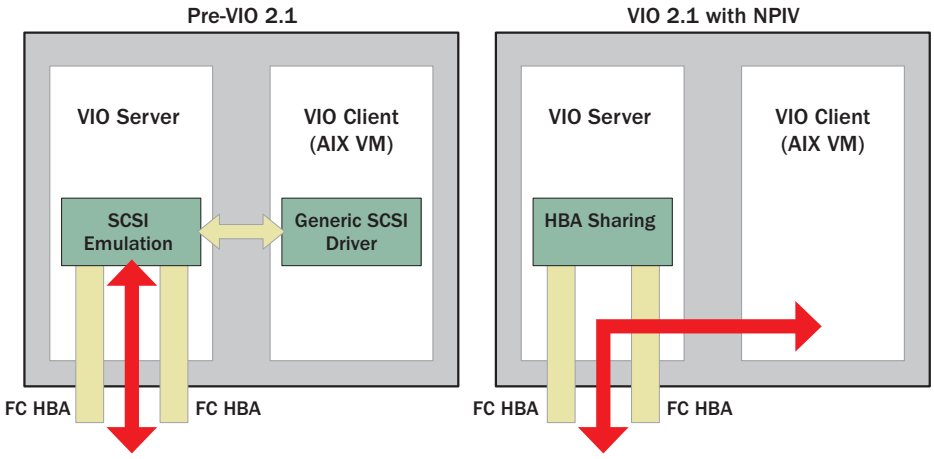

Figure 38. Before and after IBM AIX/VOI 2.1.

Pre-NPIV implementations of VIO, shown on the left in Figure 38, performed SCSI I/O through generic SCSI drivers in the VM (the VIO client) in an AIX Logical Partition (LPAR). The VIO server in another LPAR has actual control of the Fibre Channel adapters and provides SCSI emulation to all VIO clients. With VIO 2.1 and later versions, the VIO client performs I/O directly via NPIV to the Fibre Channel HBA through a virtual HBA, and the VIO server simply controls access to HBAs installed in the system, shown on the right.

The use of NPIV significantly reduces the complexity of the I/O abstraction layer. I/O is therefore less of a bottleneck and allows for more LPARs on each AIX hypervisor platform. More LPARs (VMs or VIO clients) means better consolidation ratios and the potential to save capital expenses on hypervisor platforms. I/O utilization per Fibre Channel HBA increases, perhaps necessitating the addition of more FC adapters to accommodate the increased workload. This in turn translates to higher traffic levels and more IOPS per HBA.

As consolidation of Unix hosts progresses, expect to see much higher activity at the edge of the fabric. As a result you will need to monitor the fabric much more carefully to avoid both traffic and frame congestion. It is also much more likely that the hypervisors themselves will become substantial bottlenecks.

Design Guidelines

- With the higher levels of I/O potentially occurring at each edge port in the fabric, you must ensure that there is sufficient bandwidth and paths across the fabric to accommodate the load. Consider a lot of trunked ISLs and lower subscription ratios on the ISLs, if at all possible. Remember that many flows are partially hidden due to the increased use of NPIV.
- Frame congestion is also a greater possibility. Many of the VMs may still be in clusters and may require careful configuration. Spread out the LUNs across a lot of storage ports.
- Separate the hypervisors on separate directors and, certainly, keep them separate from storage ports. This allows you to very easily apply controls through Brocade Fabric Watch classes without affecting storage.
- Determine what latencies are tolerable to both storage and hosts (VMs and storage), and consider setting Brocade FOS thresholds accordingly.
- Port Fencing is a powerful tool. Once many applications are running in VMs on a single physical platform, take care to ensure that Port Fencing does not disable ports too quickly.

Monitoring

- Bottleneck Detection becomes very important here. Use it to monitor latencies on both the hypervisor and storage ports to identify high latencies as soon as you can. Address the latencies as soon as possible.
- Brocade Fabric Watch is essential in early notification of potential issues in the fabric. Given the much higher concentration of I/O due to the server consolidation, you should closely monitor traffic levels. There is a continuing integration between Brocade Fabric Watch, Advanced Performance Monitoring, Bottleneck Detection, and Port Fencing that you should exploit to the fullest.
- Monitor the Class 3 frame discards (C3TX_TO) through Brocade Fabric Watch as well. They are a strong indication of high-latency devices.

SCALABILITY AND PERFORMANCE

Brocade products are designed with scalability in mind, knowing that most installations will continue to expand and that growth is supported with very few restrictions. However, you should follow the same basic principles outlined in previous sections as the network grows. Evaluate the impact on topology, data flow, workload, performance, and perhaps most importantly, redundancy and resiliency of the entire fabric any time one of the following actions is performed:

- Adding or removing Initiators:
  - Changes in workload
  - Changes in provisioning
- Adding or removing storage:
  - Changes in provisioning
- Adding or removing switches
- Adding or removing ISLs
- Virtualization (workload and storage) strategies and deployments

If these design best practices are followed when the network is deployed, then small incremental changes should not adversely impact the availability and performance of the network. However, if changes are ongoing and the fabric is not properly evaluated and updated, then performance and availability can be jeopardized. Some key points to cover when looking at the current status of a production FC network are these:

Reviewing redundancy and resiliency:

- Are there at least two physically independent paths between each source and destination pair?
- Are there two redundant fabrics?
- Does each host connect to two different edge switches?
- Are edge switches connected to at least two different core switches?
- Are inter-switch connections composed of two trunks of at least two ISLs?
- Does each storage device connect to at least two different edge switches or separate port blades?
- Are storage ports provisioned such that every host has at least two ports through which it can access LUNs?
- Are redundant power supplies attached to different power sources?
- Are zoning and security policies configured to allow for patch/device failover?

Reviewing performance requirements:

- Host-to-storage port fan-in/out ratios
- Oversubscription ratios:
  - Host to ISL
  - Edge switch to core switch
  - Storage to ISL
- Size of trunks
- Routing policy and currently assigned routes; evaluate actual utilization for potential imbalances

Watching for latencies such as these:

- Poor storage performance
- Overloaded hosts or applications
- Distance issues, particularly changes in usage (such as adding mirroring or too much workload)
- Deal with latencies immediately; they can have a profound impact on the fabric.

In summary, although Brocade SANs are designed to allow for any-to-any connectivity, and they support provision-anywhere implementations, these practices can have an adverse impact on the performance and availability of the SAN if left unchecked. As detailed above, the network needs to be monitored for changes and routinely evaluated for how well it meets desired redundancy and resiliency requirements.

SUPPORTABILITY

Supportability is a critical part of deploying a SAN. Follow the guidelines below to ensure that the data needed to diagnose fabric behavior or problems have been collected. While not all of these items are necessary, they are all pieces in the puzzle. You can never know which piece will be needed, so having all of the pieces available is best.

- **Configure Brocade Fabric Watch monitoring:** Leverage Brocade Fabric Watch to implement proactive monitoring of errors and warnings such as CRC errors, loss of synchronization, and high-bandwidth utilization.
- **Configure syslog forwarding:** By keeping historical log messages and having all switch messages sent to one centralized syslog server, troubleshooting can be expedited and simplified. Forwarding switch error messages to one centralized syslog server and keeping historical log messages enables faster and more effective troubleshooting and provides simple monitoring functionality.
- **Back up switch configurations:** Back up switch configurations on a regular basis to so that you can restore switch configuration in case a switch has to be swapped out or to provide change monitoring functionality.
- **Follow Brocade best practices in the LAN infrastructure:** Brocade best practices in the LAN infrastructure recommend a setup of different physical LAN broadcast segments, for example, by placing IP routers between segments or configuring different VLANs for the management interfaces of two fabric switches
- **Enable audit functionality:** To provide audit functionality for the SAN, keep track of which administrator made which changes, usage of multiple user accounts (or RADIUS), and configuration of change tracking or audit functionality (along with use of errorlog/syslog forwarding).
- **Configure multiple user accounts (LDAP/OpenLDAP or RADIUS):** Make mandatory use of personalized user accounts part of the IT/SAN security policy, so that user actions can be tracked. Also, restrict access by assigning specific user roles to individual users.
- **Establish a test bed:** Set up a test bed to test new applications, firmware upgrades, driver functionality, and scripts to avoid missteps in a production environment. Validate functionality and stability with rigorous testing in a test environment before deploying into the production environment.

- **Implement serial console server:** Implement serial remote access so that switches can be managed even when there are network issues or problems during switch boot or firmware upgrades.
- **Use aliases:** Use "aliases," which give switch ports and devices meaningful names. Using aliases to give devices meaningful names can lead to faster troubleshooting.
- **Configure supportftp:** Configure supportftp for automatic file transfers. The parameters set by this command are used by supportSave and traceDump.
- **Configure ntp server:** To keep a consistent and accurate date and time on all the switches, configure switches to use an external time server.

Firmware Upgrade Considerations

Both fixed-port and modular switches support hot code load for firmware upgrades.

- Disruptive versus non-disruptive upgrades:
- Directors versus switches
- Simultaneous upgrades on neighboring switches
- Standard FC ports versus application and special-feature ports
- Review the Brocade Fabric OS Release Notes for the following:
- Upgrade path
- Changes to feature support
- Changes to backward compatibility
- Known issues and defects
- Consider a separate AG firmware upgrade strategy. Brocade Access Gateways have no fundamental requirement to be at the same firmware release level as Brocade FOS. Upgrading only directors and switches minimizes the infrastructure changes required during an upgrade cycle.

NPIV AND THE BROCADE ACCESS GATEWAY

One of the main limits to Fibre Channel scalability is the maximum number of domains (individual physical or virtual switches) in a fabric. Keeping the number of domains low reduces much of the overhead typically attributed to SAN fabrics. Small-domain-count fabrics are more reliable, perform better, and are easier to manage. You can reduce overhead by doing the following:

- Reducing inter-switch zone transfers
- Reducing name server synchronization
- Reducing RSCN processing

The main reason for using AG mode is scalability. Given that embedded switches are smaller-port-count switches, an environment with a lot of blade servers with embedded switches can easily start to encroach on the stated limits on total domain count. Putting these switches into AG mode means they will not be consuming domain. The downside to AG mode has been the functionality (or feature set) available, although AG continues to expand its functionality with each release. Though there are some scenarios with a clear-cut answer for AG mode, generally it is an evaluation of the SAN environment and the desired functionality that determines if AG is a design option for the environment. In a fabric with lots of legacy devices, identifying and isolating misbehaving devices is easier to do in a full-fabric environment.

Last, for configurations with hosts and targets on the same AG, the traffic does need to go through the fabric switch, but it is handled within the local switch and does not need to traverse to another switch in the fabric and then back again. The theoretical domain limit in a single fabric is 239, but most fabrics are typically limited to a

much smaller number (56 is recommended in Brocade fabrics). The domain count limit typically comes into play when a large number of small-port-count switches are deployed. Large-bladed server deployments, for example, can easily push the domain count up over recommended limits when embedded blade switches are part of the implementation. FC switches in blade server enclosures typically represent fewer than 32 ports.

NPIV was originally developed to provide access to Fibre Channel devices from IBM mainframes and to improve the efficiency of mainframe I/O for virtualized environments. NPIV is part of the Fibre Channel standard and has been put to use in many open systems storage deployments. Brocade switches and directors as well as the Brocade Access Gateway support NPIV.

NPIV allows for many flows (connections) to share a single physical link. Figure 39 illustrates a single platform that supports flows from separate VMs through a single upstream link to a fabric via a shared HBA.

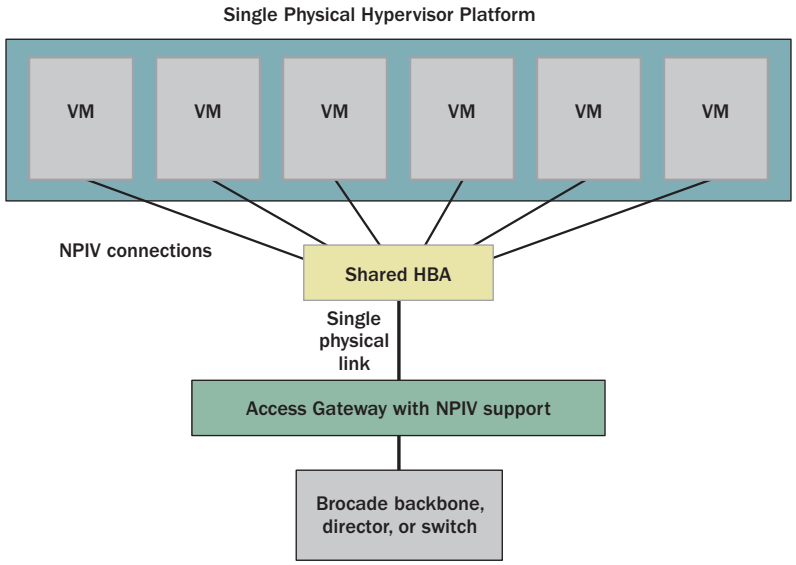

Figure 39. VMs supported on a single link to a fabric via NPIV.

A device or switch connecting to another switch via an NPIV-enabled port does not require a domain ID, does not do any zoning, and behaves much more like an end device (or group of devices) than a switch. The Brocade Access Gateway was originally designed to reduce domain ID proliferation with the introduction of embedded blade switches, which use low-port-count switches that reside in blade server chassis. In most environments, these embedded switches are deployed in large quantities, which not only lead to high-domain-count fabrics, but also increases switch administration overhead. The Brocade Access Gateway eliminates or reduces both of these issues and is supported on all Brocade embedded switches and some fixed-port switch platforms. The Brocade Access Gateway connects initiators such as host HBAs on its "downstream" F_Ports to one or more fabrics via "upstream" N_Ports.

Benefits of the Brocade Access Gateway

- **Scalability:** You can add many Access Gateways to a fabric without increasing the domain count. A major scalability constraint is avoided when small-port-count switches or embedded switches are part of an infrastructure. Registered State Change Notifications (RSCNs) are also greatly reduced—only those that are related to the initiators on the downstream Access Gateway ports are passed on through to the fabric. Since it is essentially a device, the Access Gateway can connect to more than one fabric from its upstream ports. Brocade Access Gateways can be cascaded to reduce the number of fabric connections required to support a given workload or traffic level from the attached hosts.

- **Error isolation and management:** Most initiator errors are not propagated through to the fabric. Disconnecting an upstream port, for example, does not cause a fabric rebuild. Most management activities on the Brocade Access Gateway are also isolated from the fabric. One possible scenario is server administrators managing the Access Gateways and storage administrators simply providing LUNs and zoning support for the servers using NPIV.
- **Increased resiliency:** The Brocade Access Gateway supports F_Port Trunking, which increases the resiliency of connections into the fabric. Losing a trunk member simply reduces the bandwidth of the upstream trunk. While a few frames may be lost, no host connections are affected.
- **Other:** Hosts or HBAs can be configured to automatically fail over to another upstream link, should the one they are using fail. The Brocade Access Gateway also implements many advanced features such as Adaptive Networking services (a Brocade FOS feature), Trunking, hot code load, and Brocade Fabric Watch. Brocade FOS v7.1 adds ClearLink D_Port support for Gen 5 Fibre Channel platforms, Credit Recovery, and Forward Error Correction.

Constraints

The advantages of the Brocade Access Gateway are compelling, but there are constraints:

- Although benefits are much more obvious for servers, the Brocade Access Gateway supports storage devices, but the traffic must flow through the fabric, which has its own limitations.
- There are a maximum number of 254 NPIV connections per upstream port.
- The number of Brocade Access Gateways per switch is limited only by what the fabric switches can support.

The primary factors are:

- ·The total number of devices that attach to the fabric through the Access Gateways
- The number of devices per Access Gateway N_Port
- The total number of devices attached to the switch and fabric

See the Brocade Scalability Guidelines for details.

- The number of fabrics to which a single Brocade Access Gateway can be connected is limited to the number of N_Ports on that Access Gateway. In general, most deployments require a single Access Gateway connection to only one or two fabrics. Note that the ability to connect different upstream ports to different fabrics does not reduce the requirement for redundancy. All attached servers should have dual paths to their storage through different fabrics via separate Access Gateways.

Design Guidelines

Use the Brocade Access Gateway when you deploy bladed environments or have a lot of low-port-count switches and when you need to connect different servers in different fabrics from a single-bladed enclosure. The Access Gateway can be very valuable when you want to separate the management of blade enclosures so that the enclosure is completely managed by server administrators, and the fabric is handled by storage administrators. Management separation is provided through the NPIV connection, which allows the Access Gateway to be managed separately by tools such as integrated blade server enclosure management tools without any adverse effects on the fabric.

Monitoring

Monitoring is somewhat difficult for NPIV flows. Traditional SAN monitoring has been based at the port level where hosts are connected. Multiple flows across ISLs and IFLs and into storage ports are common, but multiple host behaviors into initiators are a relatively new concept. The Brocade Access Gateway has been enhanced to include many features found in the standard version of Brocade FOS, such as Port Fencing, device security policies, and Bottleneck Detection.

Maintenance

There is usually no need to keep the Brocade Access Gateway firmware levels synchronized with the firmware levels deployed in the fabrics to which it is connected (and Brocade supports connections from other vendors' NPIV-enabled devices, where firmware synchronization is impossible). This can be significant for very large fabrics with many devices, including many Access Gateways. The version of Brocade FOS running on fabric switches can be upgraded at one time and the Access Gateways at another time, which greatly reduces the amount of change required to the infrastructure during a single maintenance window.

With the advent of Brocade FOS v7.0, end-to-end monitoring is now available for the Brocade Access Gateway. Brocade FOS v7.1 adds ClearLink D_Port support for Gen 5 Fibre Channel platforms, Credit Recovery, and Forward Error Correction.

See the *Brocade Fabric OS Release Notes* to determine if a synchronized Brocade FOS upgrade of Brocade Access Gateway devices is required.

BACKUP AND RESTORE

Backup and restore is part of an overall Disaster Recovery strategy, which itself is dependent on the criticality of data being backed up. In addition to storage consolidation, data backups are still a primary driver for a SAN-based infrastructure. This is commonly known as LAN-free backup, leveraging high-speed Fibre Channel for transport.

Note: *Since tape drives are streaming devices, it is important to determine and maintain the optimal transfer rate. Contact your tape drive vendor if this information is not available.*

The key factors for backup and restore include the following:

- Restoring backup data successfully is the most critical aspect of the backup/recovery process. In addition to ensuring business continuity in the event of a man-made or natural disaster, it is also a regulatory compliance requirement.
- Backups must be completed most, if not all, of the time.
- You should leverage backup reports so that administrators can keep track of tape media utilization and drive statistics as well as errors.
- If tapes are kept offsite for storage and Disaster Recovery, encrypt the data for security purposes.

Verify whether your industry requires data on tapes to be encrypted. Brocade offers tape and disk encryption solutions for data at rest.

Create a process and document procedures to validate backups periodically. Back up not only application data, but also include switch configurations to ensure that in the event of a switch failure a new switch can quickly be configured. Use Brocade SAN Health, Brocade Network Advisor, or the Brocade FOS CLI to capture switch configurations.

Determining SAN Bandwidth for Backups

At a minimum, available bandwidth in the fabric should be able to support applications and backup throughput. For example, in an edge-core-edge topology, the ISL paths from the storage-core tape and host-core tape should be able to support total throughput of all active tape drives and all applications without congestion. As shown in Figure 40, these paths should be redundant so that the failure of an ISL will not cause congestion in the fabric, impacting application or backup performance.

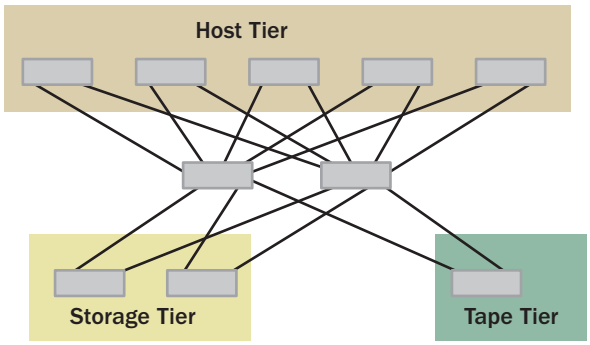

Figure 40. The same edge-core-edge tiered topology.

The key drivers for data recovery include the following:

- How quickly access to data is restored, called the Recovery Time Objective (RTO)
- The point in time in which the last valid data transaction was captured, called the Recovery Point Objective (RPO)
- Where the recovered data is located

Improving the Backup Infrastructure

Determine if the existing backup infrastructure can support expanding SANs driven by data growth:

- Look at the backup schedule and how long it takes to complete the backup, to see if there are better time periods to run the job, or schedule to a different library for faster completion.
- Use tape multiplexing or compression.

If budgets permit, other options to improve backups to meet business objectives include the following:

- Add additional drives or libraries.
- Deploy a deduplication appliance.
- Use Virtual Tape Libraries (VTLs).

From a SAN perspective, consider the following:

- Add additional ISLs, or break down existing trunks into no more than two ports in the trunk to create TI Zones. This minimizes the impact of backup traffic on other application traffic.
- Make sure that there are redundant paths to the backup tier (see the section on Device Placement for details).
- For Brocade DCX/DCX-4S Backbone chassis with open slots in the core, add a high-density port blade, such as the Brocade FC8-64, to expand the backup tier and add additional backup devices.

To reduce the time to recover from a backup, implement a two-tier disk-tape system with incremental backup to disk and migration to tape in off-hours and full backups only during downtime, such as on weekends. Another option is to implement a Continuous Data Protection (CDP) system, in which after a full backup only changed files or disk blocks are backed up. This provides the ability to restore at a granular level.

For a detailed discussion of backup and recovery concepts and issues, see Strategies for Data Protection, by Tom Clark, on Brocade Bookshelf (www.brocade.com/bookshelf).

STORAGE

Storage arrays have evolved significantly over the last few years. Performance has increased, capacities have exploded, and more LUNs are supported than ever before. The performance and capacity of low-end arrays have also improved. New features include the following:

- Some arrays time out and reset their ports if they do not receive acknowledgements from the connected host after specific intervals.
- New behaviors include using in-band Fibre Channel for control purposes, which can put extra stress on FC port buffer usage.

Note that storage array performance can degrade over time, which can be attributed to factors such as these:

- Misconfigured LUNs can impact performance.
- Provisioning strategies can favor capacity over usage. An example of this might be a policy that dictates the number of terabytes allocated per storage port. Applications accessing the LUNs can overload the array capacity in order to service the requests.

Fixing degraded array performance is never easy. It usually involves some data migration or array reconfiguration. Bottleneck Detection can be used to detect these conditions early, and changes can be implemented before performance degradation becomes chronic.

Design Guidelines

- Be careful if you deploy mixed arrays with different performance characteristics. Experience has shown that it is very easy for a Tier 3 storage array, depending on how it is used, to impact the performance of arrays in the same fabric. Troubleshooting in these situations is very difficult.
- Control the number of LUNs behind each storage port based on the type of usage they will receive.
- Check on any special short-frame traffic to avoid frame congestion at array ports. It may be necessary to increase the number of buffers at the array port to accommodate the extra control traffic.
- Use advance Brocade FOS threshold timers to monitor hosts and storage arrays to ensure that array ports do not reset due to a high-latency host, and thus do not adversely impact other connected hosts.

Monitoring

- Bottleneck Detection is indispensable; many high-latency array ports can be identified and their performance problems addressed before issues come to the attention of the server administrator.
- Use Brocade Fabric Watch to monitor Class 3 frame discards due to TX timeout so that severe latencies on storage array ports can be identified.

Storage Virtualization

Storage virtualization enables LUNs accessed by servers to be abstracted from the physical storage (typically storage arrays) on which they actually reside. (These are not the same as traditional storage array LUN allocations, which can also be viewed as a form of virtualization.) Virtualized LUNs that are disassociated from their actual storage allow for more flexible storage provisioning processes. Performance may also improve, as the virtual LUNs can be striped across multiple storage arrays.

There are two general types of storage virtualization: one uses an external controller (called in-line virtualization), and in the other, the virtualization occurs inside a storage array. In-line solutions are slightly more flexible, because they can use physical storage from a variety of sources and vendors.

Figure 41 shows a typical implementation of an in-line virtualized storage solution. The host or VM accesses storage via a storage controller (shown on top) through the storage network. The orange arrows indicate data access to and from the storage controller. The storage controller typically controls all access to the physical storage, shown on the right (and indicated by the blue arrows). This creates a very flexible storage solution, because logical LUNs can be striped across several physical arrays to improve performance, and logical LUNs can be manipulated completely transparently to the host or VM.

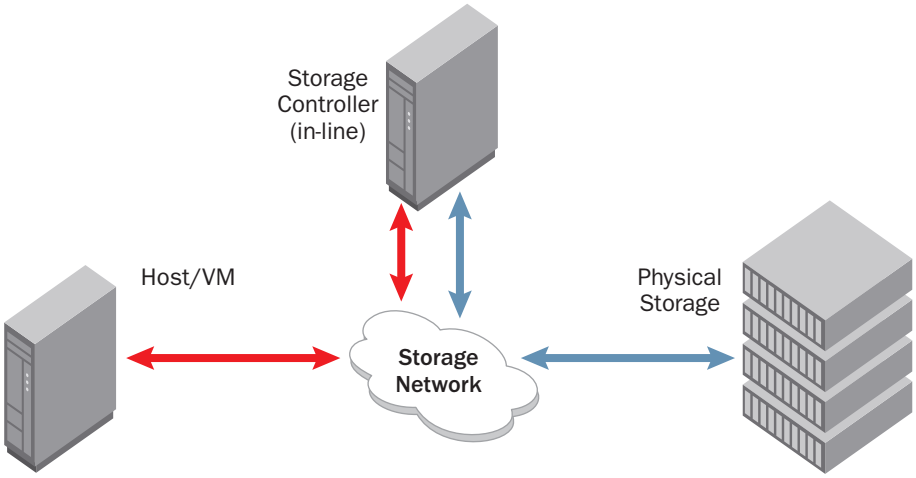

Figure 41. Typical implementation of an in-line virtualization storage solution.

The major benefit of this type of storage virtualization is that storage can now be provisioned in units of capacity (500 gigabytes or a terabyte) rather than physical LUNs. This is a first step toward viewing storage as a service instead of as physical units. VM provisioning now becomes less complex and easier to automate. Look into products such as IBM SAN Volume Controller, Hitachi Data Systems Universal Storage Platform, EMC Invista, and HP SAN Virtualization Services Platform (SVSP) for information about how these products work.

Design Guidelines

- Each storage controller in an in-line solution serves as both an initiator and a target.
- ISL utilization increases with in-line virtualized storage. Make sure that you have enough ISL bandwidth to handle the increased load.
- There is also the possibility that the in-line storage heads will communicate through Fibre Channel or generate many more SCSI control frames to manage their attached storage, which can contribute to frame congestion. You may need to increase the number of buffers at the ports that connect to the storage controller to accommodate this behavior.
- It is much more difficult to determine initiators and targets with in-line virtualized storage. Since they are on the same switch, be careful about deploying tools such as Port Fencing.

Monitoring

- Bottleneck Detection is very useful in determining latencies associated with virtualized storage.
- Brocade FOS features such as Advanced Performance Monitoring and Top Talkers are helpful in looking at high-traffic-usage flows.

SECURITY

There are many components to SAN security in relation to SAN design, and the decision to use them is greatly dependent on installation requirements rather than network functionality or performance. One clear exception is the zoning feature used to control device communication. The proper use of zoning is key to fabric functionality, performance, and stability, especially in larger networks. Other security-related features are largely mechanisms for limiting access and preventing attacks on the network (and are mandated by regulatory requirements), and they are not required for normal fabric operation.

Zoning: Controlling Device Communication

The SAN is primarily responsible for the flow of data between devices. Managing this device communication is of utmost importance for the effective, efficient, and also secure use of the storage network. Brocade Zoning plays a key role in the management of device communication. Zoning is used to specify the devices in the fabric that should be allowed to communicate with each other. If zoning is enforced, then devices that are not in the same zone cannot communicate.

In addition, zoning provides protection from disruption in the fabric. Changes in the fabric result in notifications (RSCNs) being sent to switches and devices in the fabric. Zoning puts bounds on the scope of RSCN delivery by limiting their delivery to devices when there is a change within their zone. (This also reduces the processing overhead on the switch by reducing the number of RSCNs being delivered.) Thus, only devices in the zones impacted by the change are disrupted. Based on this fact, the best practice is to create zones with one initiator and one target with which it communicates, so that changes to initiators do not impact other initiators or other targets, and disruptions are minimized (one initiator and one target device per zone). In addition, the default zone setting (what happens when zoning is disabled) should be set to No Access, which means that devices are isolated when zoning is disabled.

Zones can be defined by either switch port or device World Wide Name (WWN). While it takes a bit more effort to use WWNs in zones, it provides greater flexibility; if necessary, a device can be moved to anywhere in the fabric and maintain valid zone membership.

Zone Management: Dynamic Fabric Provisioning (DFP)

The Brocade Gen 5 Fibre Channel SAN platforms provide an integrated switch and HBA solution that enables customers to dynamically provision switch-generated virtual WWNs and create a fabric-wide zone database prior to acquiring and connecting any Brocade HBAs to the switch. DFP enables SAN administrators to pre-provision services like zoning, QoS, Device Connection Control (DCC), or any services that require port-level authentication prior to servers arriving in the fabric. This enables a more secure and flexible zoning scheme, since the fabric assigns the WWN to use. The FA-WWN can be user-generated or fabric-assigned (FA-WWN). When an HBA is replaced or a server is upgraded, zoning and LUN mapping does not have to be changed, since the new HBA is assigned the same FA-WWN as before. DFP is supported on both switches with or without the Brocade Access Gateway support. The switch automatically prevents assignment of duplicate WWNs by cross-referencing the Name Server database, but the SAN Administrator has the ultimate responsibility to prevent duplicates from being created when it is user-assigned.

Zone Management: Duplicate WWNs

In a virtual environment like VMware or HPs Virtual Connect, it is possible to encounter duplicate WWNs in the fabric. This impacts the switch response to fabric services requests like “get port WWN,” resulting in unpredictable behavior. The fabric’s handling of duplicate WWNs is not meant to be an intrusion detection tool but a recovery mechanism. Prior to Brocade FOS v7.0, when a duplicate entry is detected, a warning message is sent to the RAS log, but no effort is made to prevent the login of the second entry.

Starting with Brocade FOS v7.0, handling of duplicate WWNs is as follows:

- Same switch: The choice of which device stays in the fabric is configurable (default is to retain existing device)
- Local and remote switches: Remove both entries

Zoning recommendations include the following:

- Always enable zoning.
- Create zones with only one initiator (shown in Figure 42) and target, if possible.
- Define zones using device WWPNs (World Wide Port Names).
- Default zoning should be set to No Access.
- Use FA-WWN if supported by Brocade FOS (v7.0 or later) and Brocade HBA driver (3.0 or later).
- Delete all FA-PWWNs (Fabric-Assigned Port World Wide Names) from the switch whose configuration is being replaced before you upload or download a modified configuration.
- Follow vendor guidelines for preventing the generation of duplicate WWNs in a virtual environment.

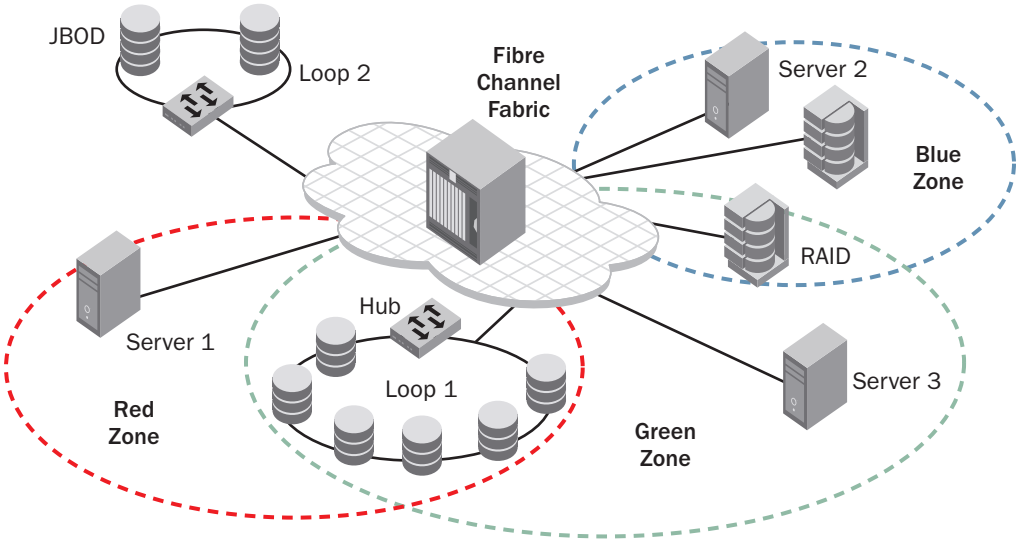

Figure 42. Example of single initiator zones.

Role-Based Access Controls (RBACs)

One way to provide limited accessibility to the fabric is through user roles. Brocade FOS has predefined user roles, each of which has access to a subset of the CLI commands. These are known as Role-Based Access Controls (RBAC), and they are associated with the user login credentials.

Access Control Lists (ACLs)

Access Control Lists are used to provide network security via policy sets. Brocade FOS provides several ACL policies including a Switch Connection Control (SCC) policy, a Device Connection Control (DCC) policy, a Fabric Configuration Server (FCS) policy, an IP Filter, and others. The following subsections briefly describe each policy and provide basic guidelines. A more in-depth discussion of ACLs can be found in the Brocade Fabric OS Administrator's Guide.

SCC Policy

The SCC policy restricts the fabric elements (FC switches) that can join the fabric. Only switches specified in the policy are allowed to join the fabric. All other switches will fail authentication if they attempt to connect to the fabric, resulting in the respective E_Ports being segmented due to the security violation.

Use the SCC policy in environments where there is a need for strict control of fabric members. Since the SCC policy can prevent switches from participating in a fabric, it is important to regularly review and properly maintain the SCC ACL.

DCC Policy

The DCC policy restricts the devices that can attach to a single FC Port. The policy specifies the FC port and one or more WWNs allowed to connect to the port. The DCC policy set comprises all of the DCC policies defined for individual FC ports. {Note that not every FC port has to have a DCC policy, and only ports with a DCC policy in the active policy set enforce access controls.} A port that is present in the active DCC policy set will allow only WWNs in its respective DCC policy to connect and join the fabric. All other devices will fail authentication when attempting to connect to the fabric, resulting in the respective F_Ports being disabled due to the security violation.

Use the DCC policy in environments where there is a need for strict control of fabric members. Since the DCC policy can prevent devices from participating in a fabric, it is important to regularly review and properly maintain the DCC policy set.

FCS Policy

Use the FCS policy to restrict the source of fabric-wide settings to one FC switch. The policy contains the WWN of one or more switches, and the first WWN (that is online) in the list is the primary FCS. If the FCS policy is active, then only the primary FCS is allowed to make and/or propagate fabric-wide parameters. These parameters include zoning, security (ACL) policies databases, and other settings.

Use the FCS policy in environments where there is a need for strict control of fabric settings. As with other ACL policies, it is important to regularly review and properly maintain the FCS policy.

IP Filter

The IP Filter policy is used to restrict access through the Ethernet management ports of a switch. Only the

IP addresses listed in the IP Filter policy are permitted to perform the specified type of activity via the management ports.

The IP Filter policy should be used in environments where there is a need for strict control of fabric access. As with other ACL policies, it is important to regularly review and properly maintain the IP Filter policy.

Authentication Protocols

Brocade FOS supports both Fibre Channel Authentication Protocols (FCAPs) and Diffie-Hellman Challenge Handshake Authentication Protocols (DH-CHAPs) on E_Ports and F_Ports. Authentication protocols provide additional security during link initialization by assuring that only the desired device/device type is connecting to a given port.

Policy Database Distribution

Security Policy Database Distribution provides a mechanism for controlling the distribution of each policy on a per-switch basis. Switches can individually configure policies to either accept or reject a policy distribution from another switch in the fabric. In addition, a fabric-wide distribution policy can be defined for the SCC and DCC policies with support for strict, tolerant, and absent modes. This can be used to enforce whether or not the SCC and/or DCC policy needs to be consistent throughout the fabric.

- **Strict mode:** All updated and new policies of the type specified (SCC, DCC, or both) must be distributed to all switches in the fabric, and all switches must accept the policy distribution.
- **Tolerant mode:** All updated and new policies of the type specified (SCC, DCC, or both) are distributed to all switches (Brocade FOS v6.2.0 or later) in the fabric, but the policy does not need to be accepted.
- **Absent mode:** Updated and new policies of the type specified (SCC, DCC, or both) are not automatically distributed to other switches in the fabric; policies can still be manually distributed.

Together, the policy distribution and fabric-wide consistency settings provide a range of control on the security policies from little or no control to very strict control.

For a detailed discussion of SAN security concepts and issues, see Securing Fibre Channel Fabrics, by Roger Bouchard, on Brocade Bookshelf (www.brocade.com/bookshelf).

CAPACITY PLANNING

Gathering Requirements

The SAN project team should interview all stakeholders (IT application owners, finance, corporate facilities, IT lab administrators, storage and network administrators, and end users) who have a vested interest in the project—and this applies equally to planning for both new and updated SANs.

Application Owners

As critical stakeholders, application owners care because everyone is measured on application uptime.

Application outages are something that users notice, and they can have severe financial impact for a business. With a redundant or a resilient infrastructure, hardware outages are transparent to the user, and only SAN administrators need to pay attention. Other questions to ask are as follows:

- What is the business goal for this application? (Is it a database that multiple applications rely on for business transactions?)
- What are the availability requirements?
- Is the application latency sensitive?
- Are there peak periods of utilization or other traffic patterns?
- What are the IOPS requirements in terms of read/writes?
- What is the worst-case response time before an outage?
- Is the application running on a cluster?
- Has the application been benchmarked to determine the CPU and memory resources required?
- Is there application downtime that can be used for applying patches, software upgrades, and maintenance?
- Can the application run on a VM? If so, how many other VMs can co-exist on the same physical hardware?

The business criticality of the application will determine the SAN design and the DR strategy, including backup and recovery. If the application is mission critical, the infrastructure must be fully redundant, with no single point of failure for both mainframe or distributed open systems architectures.

Server and Storage Administrators

Once the application requirements have been defined, identify physical server and storage on which the application and data will reside to determine the overall high-level architecture of the SAN, especially if this includes existing equipment as well as new equipment.

- Gather information about the server(s) on which the applications are running (blade or rack, CPU, memory, HBA/embedded FC switch, OS level, OS patch level, HBA driver version)?
- How many HBAs are in the rack servers?
- Does each server have single or multiple port HBAs?
- SDWhat is the primary storage for the application, and is there enough storage capacity to support this application and data? What is the current cache utilization? Is there enough cache to meet required response times?

- What is the average disk drive utilization (the greater the utilization, the longer the response times)? Contact your driver vendor to identify response times based on utilization for sizing workloads.

Utilization	25%	50%	50%
Disk drive response (milliseconds)			

- What is the raid level used? This will determine available disk space and performance for the application.
- Are storage tiers used in the environment? What is the policy used for migrating data? Are different tiers used for online storage? What is the impact?
- How many FC ports are there in the array?
- Are the arrays front-ended by a storage virtualization controller? If so, what is the additional latency?
- What are the recommended fan-in and fan-out ratios for the arrays used for this application? What are the limits?
- Is there a Disaster Recovery (DR) site? If so, how is it connected (dark fiber, FCIP)?
- What is the available/required bandwidth between the intra-site for DR? Can the existing storage infrastructure support DR with the additional load?
- What tools are used for mirroring and replication (host-based or array-based)? If host-based, was the failover tested? If so, was there any impact in application uptime? If storage=based, was the failover tested? Did the LUNs appear on the active ports? Was there an impact to application uptime?

SAN Administrator: General

A SAN administrator is responsible for the day-to-day operation of the network. The SAN design must be easy to monitor, manage, and maintain. If the current SAN is being expanded, adequate performance metrics should be collected to ensure that the existing design can be expanded to address new workloads.

- Are there performance (bandwidth) or latency issues in the existing SAN?
- Are procedures in place to address redistribution of capacity when switch port utilization exceeds 75 percent?
- Is the current design two-tier (core-edge) or three-tier (edge-core-edge)?
- Is the SAN centrally managed by a tool such as IBM Tivoli Netcool or HP OpenView?
- If there is an existing SAN, how is it managed (CLI, Brocade DCFM)? Is there a separate network for SAN management?
- Are access control policies in place for change management (zoning)? Is there a zoning policy? Are there devices in the zone database that no longer exist? What type of zoning is used (port or WWN)?
- Is the current SAN a redundant configuration?
- Is there an identified server to capture logs from the fabric?
- Is the traffic equally distributed across the ISLs or the trunks?
- Is historical performance data available for initiators, targets, and ISLs?
- How many unused ports are available per switch?

SAN Administrator: Backup and Restore

Backup and restore continue to be the primary drivers for SANs. As data growth continues to increase, backup windows continue to shrink. What is often overlooked is the restore time, which for some customers can take days.

Some topics to consider for backup and restore as you plan for SAN expansion or a new design are these:

- If the backup site is local, what is the window to complete the backup? If the backup site is remote, what is the window to complete the backup? How much of the bandwidth pipe is available?
- Is there a dedicated backup server, or do other applications share the server? Is the backup SAN on a separate SAN or a shared network?
- How often are full backups completed, and how long does it take? How often are backups checked for the integrity of the backup? How often do the backups fail to complete? What are the primary reasons (link down, tape drive failure, low throughput, other)? What is the restore time for Tier 1 and 2 applications?
- In a VM environment, is there a centralized proxy backup management, or does each VM have its own backup agent?
- Is a tiered backup implemented (disk, VTL, tape)?
- Is backup validation a regulatory requirement? If so, what processes are in place to ensure compliance?

Note: *Brocade offers certification courses in Open Systems and Mainframe SAN Design and management.*

Facilities

Facility requirements are often overlooked as SANs grow due to business expansion or data center consolidation after mergers. Even when a SAN design meets application requirements, if physical plant, power, cooling, and cable infrastructure are not available, a logically designed SAN may have to be physically distributed, which can impact application performance and ongoing servicing.

Consider the following:

- Is there existing space for new SAN devices (servers, switches, and storage)? What is the physical real estate (floor space, number of racks, rack dimensions), and do the racks have internal fans for cooling?
- What is the available power (AC 120/240), and what is the in-cabinet power and plug type? Is it the same as existing types, or do you need new power supplies?
- What method of cooling is available (hot/cool aisle, other), and what is the worst-case temperature that the data center can tolerate?
- What is the cable infrastructure (OM-3, other), and are cables already installed?
- Is there a structured cable plant with patch panels, and so forth? If so, how many patch panels will the data traverse?

Finance

Once the technical specifications have been determined, a reasonable cost estimate can be calculated based on available equipment, new purchases required, manpower, and training. Financial metrics for a total cost analysis should include the following:

- Lease versus buy
- Budget for equipment
- Budget for service and support (is 24x7 required?)
- Budget for daily operation

TOOLS FOR GATHERING DATA

Brocade SAN Health

Brocade SAN Health is a free tool that allows SAN administrators to securely capture, analyze, and report comprehensive information about Brocade fabrics with switches running Brocade FOS and M-EOS operating systems and Cisco MDS fabrics running SANOS/NXOS. It can perform tasks such as:

- Taking inventory of devices, switches, firmware versions, and SAN fabrics
- Capturing and displaying historical performance data
- Comparing zoning and switch configurations against best practices
- Assessing performance statistics and error conditions
- Producing detailed reports (in Microsoft Excel) and diagrams (in Microsoft Visio)

Note: *In mainframe FICON environments, collect the Input/Output Configuration Program (IOCP) in plain-text format (build I/O configuration statements from HCD), and upload the data. Brocade SAN Health matches the IOCP against the RNID data.*

Download Brocade SAN Health and find details and instructions on how to use it at:
www.brocade.com/services-support/drivers-downloads/san-health-diagnostics/index.page

Power Calculator

Power savings is essentially a financial issue, in terms of not only operational costs but additional costs of upgrading power infrastructure due to growth. The power calculator can be downloaded from:
www.brocade.com/data-center-best-practices/competitive-information/power.page

Storage Traffic Patterns

Most storage arrays have tools for gathering port and LUN level performance data (contact the array vendor for the appropriate tool). It is recommended that gathering a week's worth of data will help in determining if the there are enough resources to accommodate the new application requirements.

The data should reflect both normal and high utilization, such as data that reflects the end of a quarter.

The metrics to collect are as follows:

- Percent of reads
- MB/s reads
- Percent of writes
- MB/s writes
- Worst-case latency (ms)
- Number of SCSI commands/second
- Cache hits
- Queue depth

Server Traffic Patterns

On the server side, there are Windows and Unix tools for collecting CPU, memory, and network utilization built into the OS. HBA vendors also provide tools to gather the following on a per-port basis:

- Percent of reads
- MB/s reads
- Percent of writes

- MB/s writes
- Worst-case latency (ms)
- HBA queue depth

This is an example of a guideline for determining the queue depth for HBAs attached to an EMC array:

Queue depth value = 8*n/h

(where n= number of members in a metavolume group of disks, where within in the disk contiguous blocks are allocated; h = number of HBAs that can see the metavolume).

If there is an embedded switch in the server, the following information should be gathered:

- Tx frames
- Rx frames
- Total throughput

If the server hosts virtual machines, similar metrics should be collected per VM. As in the storage data collection, a week's worth of data should be collected during normal and highest utilization periods.

Backup Traffic Patterns

To understand the utilization of existing backup infrastructure, collect one week's worth of data, including when full backups are conducted. A table in Appendix B provides a template for capturing the physical infrastructure for backup.

Tape Library

If an existing SAN is used for backup, run CLI commands such as portPerfShow and portStatsShow for ports connected to the tape library and use the library management utilities to collect traffic statistics to create a profile of the current environment, to determine the following:

- Low and high utilization periods
- Drives used most often
- Tape cartridges used most often
- Tape drive volume in MB/h

Backup Media Server

On the backup media server, collect CPU, memory, FC port, and Ethernet network utilization. This helps validate that the existing backup infrastructure is working as designed to meet the backup window. It can also help determine if media server performance is impacted in a VM environment. If backup performance is impacted by non-backup traffic in the fabric, use Traffic Isolation zones or increase the number of ISLs to improve performance.

Brocade Network Advisor

Brocade Network Advisor greatly simplifies daily operations while improving the performance and reliability of the overall SAN. This software management tool offers customizable dashboards, visibility into historical data, and alert notifications to help you proactively monitor and manage your SAN network. As a result, you can optimize storage resources, maximize performance, and enhance the security of storage network infrastructures.

Brocade Network Advisor provides comprehensive management of data center fabrics, including configuration, monitoring, and management of the Brocade DCX Backbone family (including Gen 5 Fibre Channel platforms), as well as Brocade routers, switches, HBAs, and Converged Network Adapters (CNAs). Brocade Network Advisor also integrates with leading storage partner data center automation solutions to provide end-to-end network visibility through frameworks such as the Storage Management Initiative-Specification (SMI-S).

SUMMARY

Once the initial discussions with key stakeholders are complete, data should be analyzed to support an optimized SAN design given business drivers, funding, and available resources. Sometimes it can be difficult to analyze the requirements from various organizations, and creating a radar chart may help to visually analyze competing requirements from internal groups (see Appendix B). If edge switch count is increasing, consider consolidating to high-density core enterprise-level platforms, which increase port density while reducing power consumption and the number of domains to manage.

APPENDIX A: IMPORTANT TABLES

The following table shows the support distance based on cable type and data rates.

Speed Name	OM1 Link Distance 62.5-µm core and 200 MHZ*km	OM2 Link Distance 50-µm core and 500 MHz*km	OM3 Link Distance 50-µm core and 2000 MHz*km	OM4 Link Distance 50-µm core and 4700 MHz*km	OS1 Link Distance 9-µm core and ~infinite MHz*km
1GFC	300	500	860	*	10,000
2GFC	150	300	500	*	10,000
4GFC	50	150	380	400	10,000
8GFC	21	50	150	190	10,000
10GFC	33	82	300	*	10,000
16GFC	15	35	100	125	10,000

LWL Optics Support

Transceiver Data Rate (Gbps)	Distance (KM)
4	4, 10, & 30
8	10, 25
10	10
16	10

APPENDIX B: MATRICES

This section provides example checklists and tables you can use to identify dominant factors, including facilities that will have an impact on the SAN design.

Current Fabrics

SAN/Fabric	# of Switches	Type of Switches	Total Ports	Domains	# of Servers	# of Storage Devices	Location	Notes
Fabric 1								
Fabric 2								
Fabric 3								
Fabric 4								
Fabric 5								

Individual Fabric Details

SAN/Fabric	Domain Number	Serial Number	Model	Speed	WWN	IP Addresses	Brocade FOS/ M-EOS Version	Notes
Switch 1								
Switch 2								
Switch 3								
Switch 4								
Switch 5								

Device Details

Servers & Storage	Vendor	Model	WWN	Alias	Zone	OS Version	Application	Fabric/ Switches	Notes
Server 1									
Server 2									
Server 3									
Storage 1									
Storage 2									
Storage 3									

Metrics and Impact on SAN Design and Performance

The following table details the metrics that need to be collected and their impact on SAN design and performance.

Metric	Source	Impact
Servers in the SAN	Estimate/Brocade SAN Health	Normal operations
Host Level Mirroring	Estimate	Distance, ISL congestion, traffic levels
Clusters (MSFT, HACMP, NetApp) • Average Number of nodes • Workload level	Estimate Estimate: High/Med/Low	In-band heartbeat, frame congestion, host fan-in, traffic isolation
Virtualization: VIO Server # of servers Consolidation ratio	Estimate Estimate Estimate	Frame congestion, edge traffic increase/port, server fan-in on target ports, device latencies
Virtualilzation: VMware # of VMware servers Consolidated ratio Shared VMFS? DRS? RDM? I/O intensive	 Estimate Estimate Yes/No Yes (%)/No Yes (%)/No High/Med/Low	Frame congestion, device latencies, and SCSI2 reservations

Consolidated SAN Snapshot

SAN Requirements Data (Complete for each SAN)	
Fabric Information	
Target # of user ports per fabric	
Target # of total ports per fabric	
Target # of switches per fabric (# switches/switch type, total switches)	
Number of fabrics	
Number of sites in environment	
Topology (core-edge, ring, mesh, other)	
Maximum hop count	
Expected growth rate (port count)	
Fabric licenses	
SAN Device Information	
Number/types of hosts and OS platforms	
Number/types of storage devices	
Number/types of tapes	
Number/types of HBAs	
Other devices (VTL/deduplication appliance)	
Total number of SAN devices per fabric	
Customer requirement for failover/redundancy, reliability of SAN (multipathing software utilized)	
Application Details	
SAN Application (Storage Consolidation, Backup and Restore, Business Continuance)	
Fabric management application(s)	
Performance	
Maximum latency (ms)	
Targeted ISL oversubscription ratio (3:1, 7:1, 15:1, other)	

Application-Specific Details

Backup/Restore Infrastruction		
Servers		
System	**OS Version, Patch Level**	**HBA Driver Version**
Server 1/HBA		
Server 2/HBA		
Server 3/HBA		
Backup Software		
Vendor	**Version**	**Patch**
FC Switch		
Vendor	**Model**	**Firmware**
Brocade		
Storage		
Vendor	**Model**	**Firmware**
Array 1		
Array 2		
Tape Libray		
Vendor	**Model**	**Firmware**
Library		

Note: *Keep a similar table for each application.*

Quantitative Analysis: Radar Maps

SAN Admin Radar Map

SAN/Storage Admin Concerns	Rank (1 is low, 10 is high)	Notes
ISL utilization	8	Is traffic balanced across ISLs during peaks?
Switch outage	1	Have there been switch outages? If so what was the cause?
Zoning policy	6	Is the zoning policy defined?
Number of switches in the fabric	10	Is the current number of switches a concern for manageability?
Scalability	6	Can the existing design scale to support additional switches, servers, and storage?
Redundancy	10	Is the existing SAN redundant for supporting a phased migration or firmware update?
Server: High availability	10	Does the cluster software fail over reliably?
Storage: High availability	10	Do the LUNs fail over reliably?
Available disk pool	6	Is there sufficient disk pool to support additional apps?
Management tools for SAN	4	Are the right management tools used for SAN management?
Application response	7	Have there been any instances of slow application response but no outage?

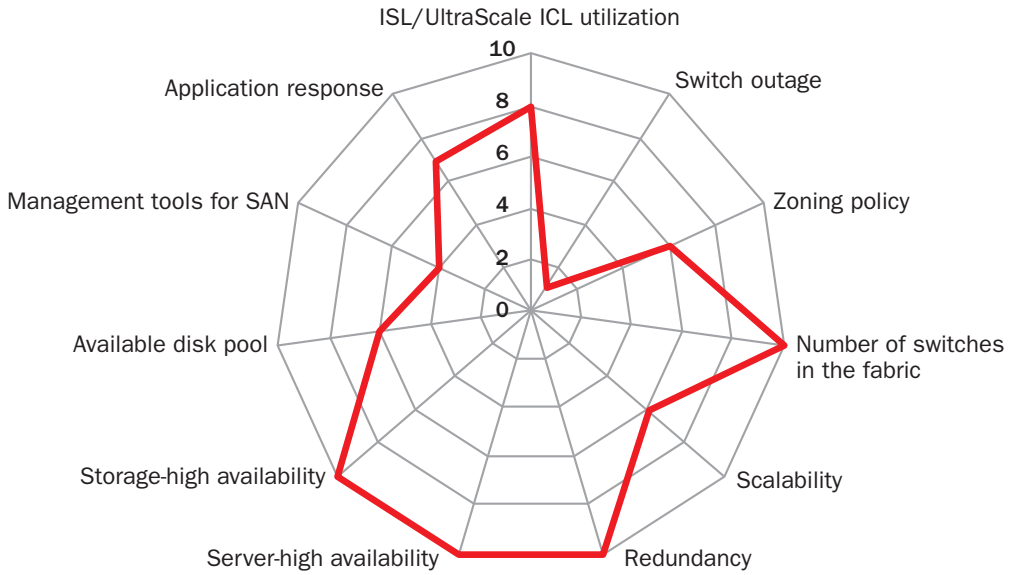

Figure 43. SAN Admin Radar Map.

Facilities Radar Map

Facility	Rank (1 is low, 10 is high)	Notes
Concern for physical real estate	8	What is the total available space for all the hardware?
Support racks	10	How many racks are needed?
Power	10	Is there adequate power?
Air conditioning	9	Is there adequate air conditioning?
Physical location	8	How important is it to have all the equipment in the same physical location or aisle?
Cable labeling	10	Are cables labeled for easy identification?
Switch labeling	10	Are switches labeled for easy identification?
Ethernet port labeling	10	Are Ethernet ports labeled for easy identification?
Patch panel labeling	10	Are patch panels labeled for easy identification?
OM-3 fiber cables used	10	Are OM-3 fiber cables in use?
Structured cabling	9	Is structured cabling in place to support SAN expansion?

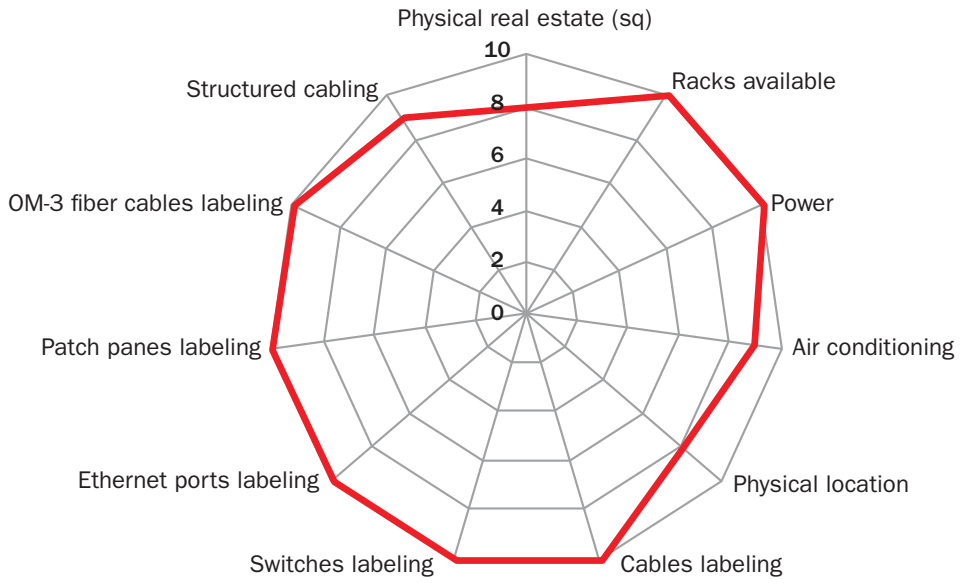

Figure 44. Facilities Radar Map.

APPENDIX C: PORT GROUPS

A port group is a group of eight ports, based on the user port number, such as 0–7, 8–15, 16–23, and up to the number of ports on the switch or port blade. Ports in a port group are usually contiguous, but they might not be. Refer to the hardware reference manual for your product for information about which ports can be used in the same port group for trunking. The ports are color-coded to indicate which can be used in the same port group for trunking (trunking port groups can be up to eight ports).

Brocade 5300 Trunk Port Groups

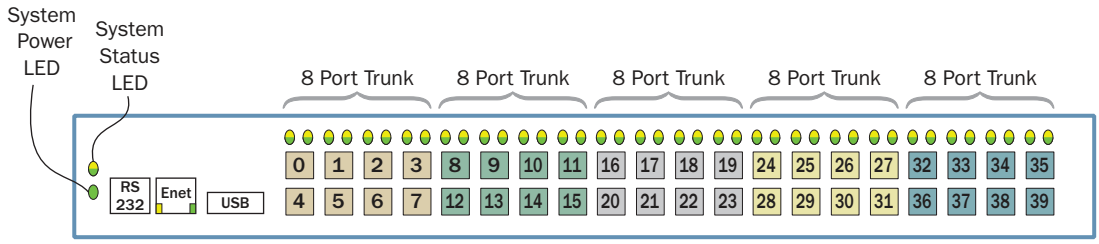
System Power LED
System Status LED
8 Port Trunk
8 Port Trunk
8 Port Trunk
8 Port Trunk
8 Port Trunk
RS 232
Enet
USB
0 1 2 3 8 9 10 11 16 17 18 19 24 25 26 27 32 33 34 35
4 5 6 7 12 13 14 15 20 21 22 23 28 29 30 31 36 37 38 39

Figure 45. Brocade 5300 Trunk Port Groups.

Brocade FC8-64 Trunk Groups

The Brocade FC8-64 port blade uses 4 Condor2 ASICs, and the figure below shows the ASIC boundaries for planning ISL trunks and end-nodes configured for local switching. Up to eight 8-port trunk groups can be created with the 64-port blade.

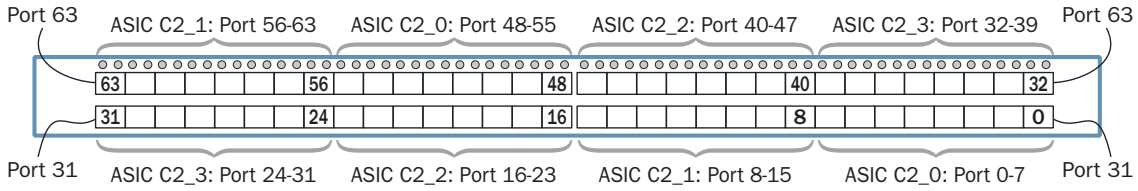

Figure 46. Brocade FC8-64 Trunk Groups.

Gen 5 Fibre Channel Platforms

The Brocade FC16-48 and FC16-32 blades for the Brocade DCX 8510 Backbone and the Brocade 6520, 6510, and 6505 Switches provide trunk groups with a maximum of 8 ports per trunk group. The trunking octet groups are in the following blade port ranges: 0-7, 8-15, 16-23, 24-31, 32-39, and 40-47. (Trunk groups 32-39 and 40-47 are obviously not applicable to FC16-32). Trunk boundary layout is on the faceplate of the blade.

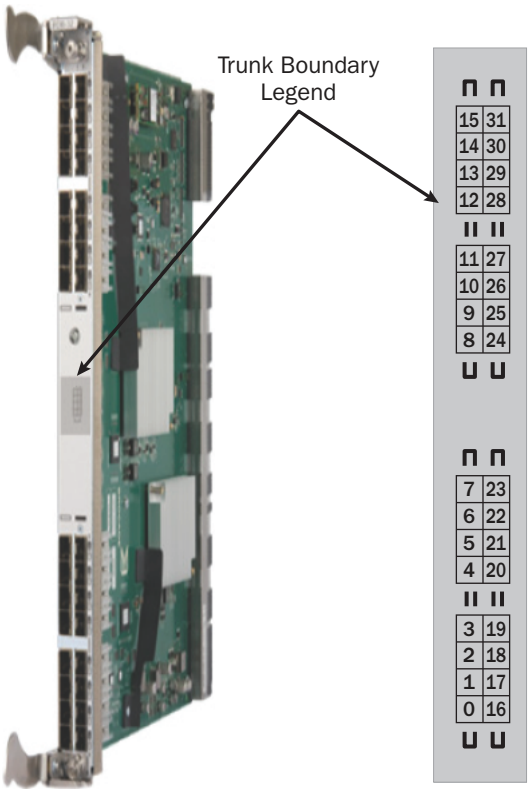

Figure 47. Brocade FC16-32 Trunk Groups.

Brocade 6520 Trunk Groups

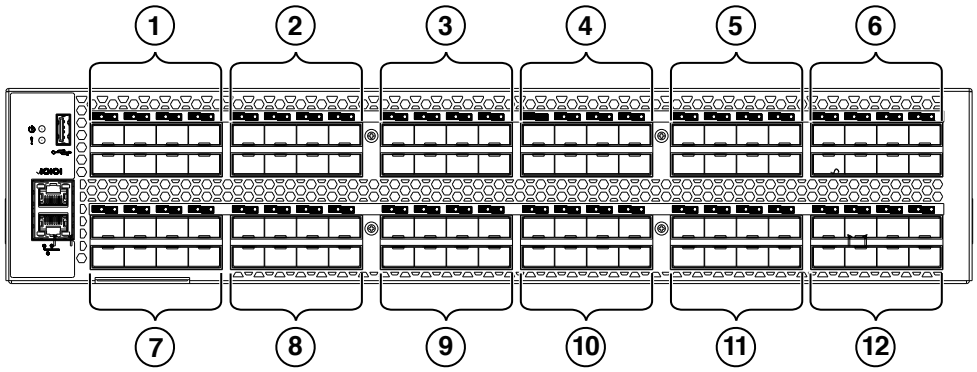

Figure 48. Brocade 6520 Front Port Groups

Brocade 6510 Trunk Groups

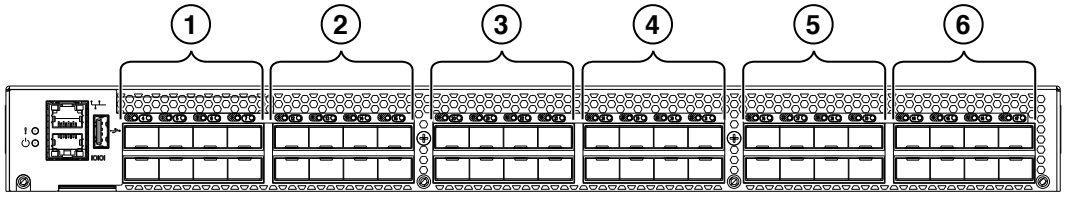

Figure 49. Brocade 6510 Front Port Groups

Brocade 6505 Trunk Groups

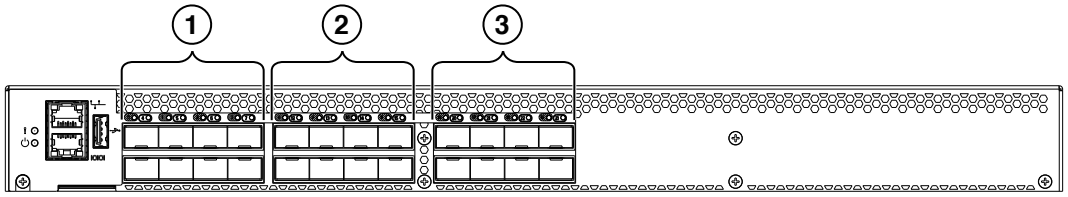

Figure 50. Brocade 6505 Front Port Groups

APPENDIX D: TERMINOLOGY

Term	Brief Description
Base switch	Base switch of an enabled virtual fabric mode switch
ClearLink Diagnostics	Diagnostics tool that allows users to automate a battery of tests to verify the integrity of all 16 Gbps transceivers in the fabric
Default switch	Default switch of an enabled virtual fabric mode switch
E_Port	A standard Fibre Channel mechanism that enables switches to network with each other
Edge Hold Time	Enables the switch to time out frames for F_Ports sooner than for E_Ports
EX_Port	A type of E_Port that connects a Fibre Channel router to an edge fabric
F_Port	A fabric port to which an N_Port is attached
FCIP	Fibre Channel over IP, which enables Fibre Channel traffic to flow over an IP link
FCR	Fibre Channel Routing, which enables multiple fabrics to share devices without having to merge the fabrics
IFL	Inter-Fabric Link, a link between fabrics in a routed topology
ISL	Inter-Switch Link, used for connecting fixed port and modular switches
Logical Switch	Logical Switch of an enabled virtual fabric mode switch
Oversubscription	A condition in which more devices might need to access a resource than that resource can fully support
Port group	A set of sequential ports that are defined (for example, ports 0–3)
QoS	Quality of Service traffic shaping feature that allows the prioritization of data traffic based on the SID/DID of each frame
Redundant	Duplication of components, including an entire fabric, to avoid a single point of failure in the network (fabrics A & B are identical)
Resilient	Ability of a fabric to recover from failure, could be in a degraded state but functional (for example, ISL failure in a trunk group)
TI Zone	Traffic Isolation Zone, which controls the flow of interswitch traffic by creating a dedicated path for traffic flowing from a specific set of source ports
Trunk	Trunking that allows a group of ISLs to merge into a single logical link, enabling traffic to be distributed dynamically at the frame level
UltraScale ICL	UltraScale Inter-Chassis Link, used for connecting modular switches without using front-end device ports
VC	Virtual channels, which create multiple logical data paths across a single physical link or connection
VF	Virtual fabrics, a suite of related features that enable customers to create a Logical Switch, a Logical Fabric, or share devices in a Brocade Fibre Channel SAN

APPENDIX E: REFERENCES

Software and Hardware Product Documentation

- *Brocade Fabric OS v7.1 Release Notes*
- *Brocade Fabric OS Administrator's Guide,* supporting Brocade Fabric OS v7.1
- *Brocade Fabric OS Command Reference Manual*, supporting Brocade Fabric OS v7.1
- *Brocade Fabric Watch Administrator's Guide,* supporting Brocade Fabric OS v7.1
- *Brocade Access Gateway Administrator's Guide,* supporting Brocade Fabric OS v7.1
- *Brocade Fabric OS Troubleshooting and Diagnostics Guide,* supporting Brocade Fabric OS v7.1
- *Hardware Reference Guides and QuickStart Guides* for backbone, director, switch, and blade platforms

Technical Briefs

www.brocade.com/sites/dotcom/data-center-best-practices/resource-center/index.page

www.brocade.com/products/all/san-backbones/product-details/dcx8510-backbone/specifications.page

Brocade Compatibility Matrix

www.brocade.com/forms/getFile?p=documents/matrices/compatibility-matrix-fos-7x-mx.pdf

Brocade Scalability Guidelines

www.brocade.com/forms/getFile?p=documents/matrices/scalability-matrix-fos-v7.0a.pdf

Brocade SAN Health

www.brocade.com/services-support/drivers-downloads/san-health-diagnostics/index.page

Brocade FOS Features

http://www.brocade.com/forms/getFile?p=documents/brochures/FabricOS_Guide.pdf

Brocade Bookshelf

- *Principles of SAN Design* (updated in 2007) by Josh Judd
- *Strategies for Data Protection* by Tom Clark
- *Securing Fibre Channel Fabrics* by Roger Bouchard (updated 2012)
- *The New Data Center* by Tom Clark

Other

www.snia.org/education/dictionary

www.vmware.com/pdf/vi3_san_design_deploy.pdf

www.vmware.com/files/pdf/vcb_best_practices.pdf

© 2013 Brocade Communications Systems, Inc. All Rights Reserved. 05/13 GA-TB-329-04

ADX, AnyIO, Brocade, Brocade Assurance, the B-wing symbol, DCX, Fabric OS, ICX, MLX, MyBrocade, OpenScript, VCS, VDX, and Vyatta are registered trademarks, and HyperEdge, The Effortless Network, and The On-Demand Data Center are trademarks of Brocade Communications Systems, Inc., in the United States and/or in other countries. Other brands, products, or service names mentioned may be trademarks of their respective owners.

Notice: This document is for informational purposes only and does not set forth any warranty, expressed or implied, concerning any equipment, equipment feature, or service offered or to be offered by Brocade. Brocade reserves the right to make changes to this document at any time, without notice, and assumes no responsibility for its use. This informational document describes features that may not be currently available. Contact a Brocade sales office for information on feature and product availability. Export of technical data contained in this document may require an export license from the United States government.